

Rényi Divergence and Kullback-Leibler Divergence

Tim van Erven

Peter Harremoës, *Member, IEEE*

Abstract—Rényi divergence is related to Rényi entropy much like Kullback-Leibler divergence is related to Shannon's entropy, and comes up in many settings. It was introduced by Rényi as a measure of information that satisfies almost the same axioms as Kullback-Leibler divergence, and depends on a parameter that is called its order. In particular, the Rényi divergence of order 1 equals the Kullback-Leibler divergence.

We review and extend the most important properties of Rényi divergence and Kullback-Leibler divergence, including convexity, continuity, limits of σ -algebras and the relation of the special order 0 to the Gaussian dichotomy and contiguity. We also extend the known equivalence between channel capacity and minimax redundancy to continuous channel inputs (for all orders), and present several other minimax results.

Index Terms—channel capacity, Kullback-Leibler divergence, minimax redundancy, Rényi divergence

I. INTRODUCTION

SHANNON entropy and Kullback-Leibler divergence (also known as information divergence or relative entropy) are perhaps the two most fundamental quantities in information theory and its applications. Because of their success, there have been many attempts to generalise these concepts, and in the literature one will find numerous entropy and divergence measures. Most of these quantities have never found any applications, and almost none of them have found an interpretation in terms of coding. The most important exceptions are the Rényi entropy and Rényi divergence [1]. Harremoës [2] and Grünwald [3, p. 649] provide an operational characterization of Rényi divergence as the number of bits by which a mixture of two codes can be compressed; and Csiszár [4] gives an operational characterization of Rényi divergence as the cut-off rate in block coding and hypothesis testing.

Rényi divergence appears as a crucial tool in proofs of convergence of minimum description length and Bayesian estimators, both in parametric and nonparametric models [5], [6], and one may recognize it implicitly in many computations throughout information theory. It is also closely related to Hellinger distance, which is commonly used in the analysis of nonparametric density estimation [7]–[9]. Rényi himself used his divergence to prove the convergence of state probabilities in a stationary Markov chain to the stationary distribution [1], and still other applications of Rényi divergence can be found, for instance, in hypothesis testing [10], in multiple source adaptation [11] and in ranking of images [12].

Although the closely related Rényi entropy is well studied [13], [14], the properties of Rényi divergence are scattered throughout the literature and have often only been established

for finite alphabets. This paper is intended as a reference document, which treats the most important properties of Rényi divergence in detail, including Kullback-Leibler divergence as a special case. Preliminary versions of the results presented here can be found in [15] and [16]. During the preparation of this paper, Shayevitz has independently published closely related work [17], [18].

A. Rényi's Information Measures

For finite alphabets, the *Rényi divergence* of positive order $\alpha \neq 1$ of a probability distribution $P = (p_1, \dots, p_n)$ from another distribution $Q = (q_1, \dots, q_n)$ is

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1} \ln \sum_{i=1}^n p_i^\alpha q_i^{1-\alpha}, \quad (1)$$

where, for $\alpha > 1$, we read $p_i^\alpha q_i^{1-\alpha}$ as $p_i^\alpha / q_i^{(\alpha-1)}$ and adopt the conventions that $\frac{0}{0} = 0$ and $\frac{x}{0} = \infty$ for $x > 0$. As described in Section II, this definition generalises to continuous spaces by replacing the probabilities by densities and the sum by an integral. If P and Q are members of the same exponential family, then their Rényi divergence can be computed using a formula by Liese and Vajda [19, p. 43], [10]. Gil [20] provides a long list of examples.

Example 1. Let Q be a probability distribution and A a set with positive probability. Let P be the conditional distribution of Q given A . Then

$$D_\alpha(P\|Q) = -\ln Q(A).$$

We observe that in this important special case the factor $\frac{1}{\alpha-1}$ in the definition of Rényi divergence has the effect that the value of $D_\alpha(P\|Q)$ does not depend on α .

The Rényi entropy

$$H_\alpha(P) = \frac{1}{1-\alpha} \ln \sum_{i=1}^n p_i^\alpha$$

can be expressed in terms of the Rényi divergence of P from the uniform distribution $U = (\frac{1}{n}, \dots, \frac{1}{n})$:

$$H_\alpha(P) = H_\alpha(U) - D_\alpha(P\|U) = \ln n - D_\alpha(P\|U).$$

As α tends to 1, the Rényi entropy tends to the Shannon entropy and the Rényi divergence tends to the Kullback-Leibler divergence, so we recover a well-known relation. The *differential Rényi entropy* of a distribution P with density p is given by

$$h_\alpha(P) = \frac{1}{1-\alpha} \ln \int (p(x))^\alpha dx$$

Tim van Erven (tim@timvanerven.nl) is with the Département de Mathématiques, Université Paris-Sud, France. Peter Harremoës (harremoës@ieee.org) is with the Copenhagen Business College, Denmark. Some of the results in this paper have previously been presented at the ISIT 2010 conference.

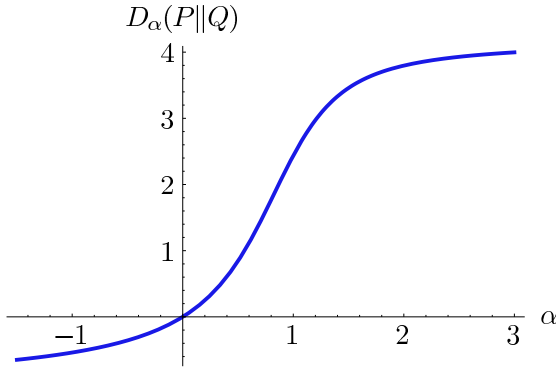


Fig. 1. Rényi divergence as a function of its order for fixed distributions

whenever this integral is defined. If P has support in an interval I of length n then

$$h_\alpha(P) = \ln n - D_\alpha(P||U_I),$$

where U_I denotes the uniform distribution on I . Thus the properties of both the Rényi entropy and the differential Rényi entropy can be deduced from the properties of Rényi divergence as long as P has compact support.

There is another way of relating Rényi entropy and Rényi divergence, in which entropy is considered as self-information. Let X denote a discrete random variable with distribution P , and let P_{diag} be the distribution of (X, X) . Then

$$H_\alpha(P) = D_{2-\alpha}(P_{\text{diag}}||P \times P).$$

For α tending to 1, the right-hand side tends to the mutual information between X and itself, and again a well-known formula is recovered.

B. Special Orders

Although one can define the Rényi divergence of any order, certain values have wider application than others. Of particular interest are the values 0, $\frac{1}{2}$, 1, 2, and ∞ .

The values 0, 1, and ∞ are *extended orders* in the sense that Rényi divergence of these orders cannot be calculated by plugging into (1). Instead, their definitions are determined by continuity in α . (See Figure 1.) This leads to defining Rényi divergence of order 1 as the Kullback-Leibler divergence. For order 0 it becomes $-\ln Q(\{i \mid p_i > 0\})$, which is closely related to absolute continuity and contiguity of the distributions P and Q (see Section III-F). For order ∞ , Rényi divergence is defined as $\ln \max_i \frac{p_i}{q_i}$. In the literature on the *minimum description length principle* in statistics, this is called the *worst-case regret* of coding with Q rather than with P [3]. The Rényi divergence of order ∞ is also related to the *separation distance*, used by Aldous and Diaconis [21] to bound the rate of convergence to the stationary distribution for certain Markov chains.

Only for $\alpha = 1/2$ is Rényi divergence symmetric in its arguments. Although not itself a metric, it is a function of the square of the Hellinger distance $\text{Hel}^2(P, Q) = \sum_{i=1}^n (\sqrt{p_i} - \sqrt{q_i})^2$ [22]:

$$D_{\frac{1}{2}}(P||Q) = -2 \ln \left(1 - \frac{\text{Hel}^2(P, Q)}{2} \right). \quad (2)$$

Similarly, for $\alpha = 2$ it satisfies

$$D_2(P||Q) = \ln(1 + \chi^2(P, Q)), \quad (3)$$

where $\chi^2(P, Q) = \sum_{i=1}^n \frac{(p_i - q_i)^2}{q_i}$ denotes the χ^2 -divergence [22]. It will be shown that Rényi divergence is nondecreasing in its order. Therefore, by $\ln t \leq t - 1$, (2) and (3) imply that

$$\begin{aligned} \text{Hel}^2(P, Q) &\leq D_{\frac{1}{2}}(P||Q) \leq D_1(P||Q) \\ &\leq D_2(P||Q) \leq \chi^2(P, Q). \end{aligned}$$

Finally, Gilardoni [23] shows that Rényi divergence of orders $\alpha \in (0, 1]$ is related to the total variation distance $V(P, Q) = \sum_{i=1}^n |q_i - p_i|$ by a generalisation of *Pinsker's inequality*:

$$\frac{\alpha}{2} V^2(P, Q) \leq D_\alpha(P||Q). \quad (4)$$

For $\alpha = 1$ this is the normal version of Pinsker's inequality, which bounds total variation distance in terms of the square root of the Kullback-Leibler divergence.

C. Outline

The rest of the paper is organized as follows. First, in Section II, we extend the definition of Rényi divergence from formula (1) to continuous spaces. One can either define Rényi divergence via an integral or via discretizations. We demonstrate that these definitions are equivalent. Then we show that Rényi divergence extends to the extended orders 0, 1 and ∞ in the same way as for finite spaces. Along the way, we also study its behaviour as a function of α . By contrast, in Section III we study various convexity and continuity properties of Rényi divergence as a function of P and Q , while α is kept fixed. Section IV contains several minimax results, and treats the connection to Chernoff information in hypothesis testing, to which many applications of Rényi divergence are related. We also discuss the equivalence of channel capacity and the minimax redundancy for all orders α . Then, in Section V, the main part of the paper is completed by an extension of Rényi divergence to negative orders. These are related to the orders $\alpha > 1$ by a negative scaling factor and a reversal of the arguments P and Q . Finally, the appendix contains a number of negative results, i.e. examples showing that properties that hold for certain other divergences are violated by Rényi divergence.

For fixed α , Rényi divergence is related to various forms of *power divergences*, which are in the well-studied class of *f-divergences* [24]. Consequently, several of the results we are presenting for fixed α in Section III are equivalent to known results about power divergences. To make this presentation self-contained we avoid the use of such connections and only use general results from measure theory.

II. DEFINITION OF RÉNYI DIVERGENCE

Let us fix the notation to be used throughout the paper. We consider (probability) measures on a measurable space $(\mathcal{X}, \mathcal{F})$. If P is a measure on $(\mathcal{X}, \mathcal{F})$, then we write $P|_{\mathcal{G}}$ for its restriction to the σ -subalgebra $\mathcal{G} \subseteq \mathcal{F}$, which may be interpreted as the marginal of P on the subset of events \mathcal{G} . A measure P is

called *absolutely continuous* with respect to another measure Q if $P(A) = 0$ whenever $Q(A) = 0$ for all events $A \in \mathcal{F}$. We will write $P \ll Q$ if P is absolutely continuous with respect to Q and $P \not\ll Q$ otherwise. Alternatively, P and Q may be *mutually singular*, denoted $P \perp Q$, which means that there exists an event $A \in \mathcal{F}$ such that $P(A) = 0$ and $Q(\mathcal{X} \setminus A) = 0$. We will assume that all (probability) measures are absolutely continuous with respect to a common σ -finite measure μ , which is arbitrary in the sense that none of our definitions or results depend on the choice of μ . As we only consider (mixtures of) a countable number of distributions, such a measure μ exists in all cases, so this is no restriction. For measures denoted by capital letters (e.g. P or Q), we will use the corresponding lower-case letters (e.g. p, q) to refer to their densities with respect to μ . And for any event $A \in \mathcal{F}$, $\mathbf{1}_A$ denotes its indicator function, which is 1 on A and 0 otherwise. Finally, we use the constant $\tau = 2\pi$ to slightly simplify some expressions, and use the natural logarithm in our definitions, such that information is measured in nats (1 bit equals $\ln 2$ nats).

We will often need to distinguish between the orders for which Rényi divergence can be defined by a generalisation of formula (1) to an integral over densities, and the other orders. This motivates the following definitions.

Definition 1. We call a (finite) real number α a *simple order* if $\alpha > 0$ and $\alpha \neq 1$. The values 0, 1, and ∞ are called *extended orders*.

A. Definition by Formula for Simple Orders

Let P and Q be two arbitrary distributions on $(\mathcal{X}, \mathcal{F})$. The formula in (1), which defines Rényi divergence for simple orders on finite sample spaces, generalises to arbitrary spaces as follows:

Definition 2 (Simple Orders). For any simple order α , the *Rényi divergence of order α* of P from Q is defined as

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \ln \int p^\alpha q^{1-\alpha} d\mu, \quad (5)$$

where, for $\alpha > 1$, we read $p^\alpha q^{1-\alpha}$ as $\frac{p^\alpha}{q^{\alpha-1}}$ and adopt the conventions that $\frac{0}{0} = 0$ and $\frac{x}{0} = \infty$ for $x > 0$.

For example, for any simple order α , the Rényi divergence of a normal distribution (with mean μ_0 and positive variance σ_0^2) from another normal distribution (with mean μ_1 and positive variance σ_1^2) is

$$D_\alpha(\mathcal{N}(\mu_0, \sigma_0^2)||\mathcal{N}(\mu_1, \sigma_1^2)) = \frac{\alpha(\mu_1 - \mu_0)^2}{2\sigma_\alpha^2} + \frac{1}{1 - \alpha} \ln \frac{\sigma_\alpha}{\sigma_0^{1-\alpha} \sigma_1^\alpha}, \quad (6)$$

provided that $\sigma_\alpha^2 = (1 - \alpha)\sigma_0^2 + \alpha\sigma_1^2 > 0$ [19, p. 45].

The interpretation of $p^\alpha q^{1-\alpha}$ in Definition 2 is such that the *Hellinger integral* $\int p^\alpha q^{1-\alpha} d\mu$ is an f -divergence [24], which ensures that the relations from the introduction to squared Hellinger distance (2), χ^2 -distance (3), and the total variation distance (4) hold in general, not just for finite sample spaces.

For simple orders, we may always change to integration with respect to P :

$$\int p^\alpha q^{1-\alpha} d\mu = \int \left(\frac{q}{p}\right)^{1-\alpha} dP,$$

which shows that our definition does not depend on the choice of dominating measure μ . In most cases it is also equivalent to integrate with respect to Q :

$$\int p^\alpha q^{1-\alpha} d\mu = \int \left(\frac{p}{q}\right)^\alpha dQ \quad (0 < \alpha < 1 \text{ or } P \ll Q).$$

However, if $\alpha > 1$ and $P \not\ll Q$, then $D_\alpha(P||Q) = \infty$, whereas the integral with respect to Q may be finite.

B. Definition via Discretization for Simple Orders

We shall repeatedly use the following result, which is a direct consequence of the Radon-Nikodým theorem [25]:

Proposition 1. Suppose $\lambda \ll \mu$ is a probability distribution, or any countably additive measure such that $\lambda(\mathcal{X}) \leq 1$. Then for any σ -subalgebra $\mathcal{G} \subseteq \mathcal{F}$

$$\frac{d\lambda|_{\mathcal{G}}}{d\mu|_{\mathcal{G}}} = \mathbf{E} \left[\frac{d\lambda}{d\mu} \middle| \mathcal{G} \right] \quad (\mu\text{-a.s.})$$

It has been argued that grouping observations together (by considering a coarser σ -algebra), should not increase our ability to distinguish between P and Q under any measure of divergence [26]. This is expressed by the *data processing inequality*, which Rényi divergence satisfies:

Theorem 1 (Data Processing Inequality). For any simple order α and any σ -subalgebra $\mathcal{G} \subseteq \mathcal{F}$

$$D_\alpha(P|_{\mathcal{G}}||Q|_{\mathcal{G}}) \leq D_\alpha(P||Q).$$

Theorem 9 below shows that the data processing inequality also holds for the extended orders.

Proof: Let \tilde{P} denote the absolutely continuous component of P with respect to Q . Then by Proposition 1 and Jensen's inequality for conditional expectations

$$\begin{aligned} \frac{1}{\alpha - 1} \ln \int \left(\frac{d\tilde{P}|_{\mathcal{G}}}{dQ|_{\mathcal{G}}} \right)^\alpha dQ &= \frac{1}{\alpha - 1} \ln \int \left(\mathbf{E} \left[\frac{d\tilde{P}}{dQ} \middle| \mathcal{G} \right] \right)^\alpha dQ \\ &\leq \frac{1}{\alpha - 1} \ln \int \mathbf{E} \left[\left(\frac{d\tilde{P}}{dQ} \right)^\alpha \middle| \mathcal{G} \right] dQ \\ &= \frac{1}{\alpha - 1} \ln \int \left(\frac{d\tilde{P}}{dQ} \right)^\alpha dQ. \end{aligned} \quad (7)$$

If $0 < \alpha < 1$, then $p^\alpha q^{1-\alpha} = 0$ if $q = 0$, so the restriction of P to \tilde{P} does not change the Rényi divergence, and hence the theorem is proved. Alternatively, suppose $\alpha > 1$. If $P \ll Q$, then $\tilde{P} = P$ and the theorem again follows from (7). If $P \not\ll Q$, then $D_\alpha(P||Q) = \infty$ and the theorem holds as well. ■

The next theorem shows that if \mathcal{X} is a continuous space, then the Rényi divergence on \mathcal{X} can be arbitrarily well

approximated by the Rényi divergence on finite partitions of \mathcal{X} . For any finite or countable partition $\mathcal{P} = \{A_1, A_2, \dots\}$ of \mathcal{X} , let $P|_{\mathcal{P}} \equiv P|_{\sigma(\mathcal{P})}$ and $Q|_{\mathcal{P}} \equiv Q|_{\sigma(\mathcal{P})}$ denote the restrictions of P and Q to the σ -algebra generated by \mathcal{P} .

Theorem 2. *For any simple order α*

$$D_\alpha(P\|Q) = \sup_{\mathcal{P}} D_\alpha(P|_{\mathcal{P}}\|Q|_{\mathcal{P}}), \quad (8)$$

where the supremum is over all finite partitions $\mathcal{P} \subseteq \mathcal{F}$.

It follows that it would be equivalent to first define Rényi divergence for finite sample spaces and then extend the definition to arbitrary sample spaces using (8).

The identity (8) also holds for the extended orders 1 and ∞ . (See Theorem 10 below.)

Proof of Theorem 2: By the data processing inequality

$$\sup_{\mathcal{P}} D_\alpha(P|_{\mathcal{P}}\|Q|_{\mathcal{P}}) \leq D_\alpha(P\|Q).$$

To show the converse inequality, consider for any $\varepsilon > 0$ a discretisation of the densities p and q into a countable number of bins

$$B_{m,n}^\varepsilon = \{x \in \mathcal{X} \mid e^{m\varepsilon} \leq p(x) < e^{(m+1)\varepsilon}, \\ e^{n\varepsilon} \leq q(x) < e^{(n+1)\varepsilon}\},$$

where $n, m \in \{-\infty, \dots, -1, 0, 1, \dots\}$. Let $\mathcal{Q}^\varepsilon = \{B_{m,n}^\varepsilon\}$ and $\mathcal{F}^\varepsilon = \sigma(\mathcal{Q}^\varepsilon) \subseteq \mathcal{F}$ be the corresponding partition and σ -algebra, and let $p_\varepsilon = dP|_{\mathcal{Q}^\varepsilon}/d\mu$ and $q_\varepsilon = dQ|_{\mathcal{Q}^\varepsilon}/d\mu$ be the densities of P and Q restricted to \mathcal{F}^ε . Then by Proposition 1

$$\frac{q_\varepsilon}{p_\varepsilon} = \frac{\mathbf{E}[q \mid \mathcal{F}^\varepsilon]}{\mathbf{E}[p \mid \mathcal{F}^\varepsilon]} \leq \frac{q}{p} e^{2\varepsilon} \quad (P\text{-a.s.})$$

It follows that

$$\frac{1}{\alpha - 1} \ln \int \left(\frac{q_\varepsilon}{p_\varepsilon}\right)^{1-\alpha} dP \geq \frac{1}{\alpha - 1} \ln \int \left(\frac{q}{p}\right)^{1-\alpha} dP - 2\varepsilon,$$

and hence the supremum over all countable partitions is large enough:

$$\sup_{\substack{\text{countable } \mathcal{Q} \\ \sigma(\mathcal{Q}) \subseteq \mathcal{F}}} D_\alpha(P|_{\mathcal{Q}}\|Q|_{\mathcal{Q}}) \geq \sup_{\varepsilon > 0} D_\alpha(P|_{\mathcal{Q}^\varepsilon}\|Q|_{\mathcal{Q}^\varepsilon}) \geq D_\alpha(P\|Q).$$

It remains to show that the supremum over finite partitions is at least as large. To this end, suppose $\mathcal{Q} = \{B_1, B_2, \dots\}$ is any countable partition and let $\mathcal{P}_n = \{B_1, \dots, B_{n-1}, \bigcup_{i \geq n} B_i\}$. Then by

$$P\left(\bigcup_{i \geq n} B_i\right)^\alpha Q\left(\bigcup_{i \geq n} B_i\right)^{1-\alpha} \geq 0 \quad (\alpha > 1),$$

$$\lim_{n \rightarrow \infty} P\left(\bigcup_{i \geq n} B_i\right)^\alpha Q\left(\bigcup_{i \geq n} B_i\right)^{1-\alpha} = 0 \quad (0 < \alpha < 1),$$

we find that

$$\begin{aligned} \lim_{n \rightarrow \infty} D_\alpha(P|_{\mathcal{P}_n}\|Q|_{\mathcal{P}_n}) &= \lim_{n \rightarrow \infty} \frac{1}{\alpha - 1} \ln \sum_{B \in \mathcal{P}_n} P(B)^\alpha Q(B)^{1-\alpha} \\ &\geq \lim_{n \rightarrow \infty} \frac{1}{\alpha - 1} \ln \sum_{i=1}^{n-1} P(B_i)^\alpha Q(B_i)^{1-\alpha} \\ &= D_\alpha(P|_{\mathcal{Q}}\|Q|_{\mathcal{Q}}), \end{aligned}$$

where the inequality holds with equality if $0 < \alpha < 1$. ■

C. Extended Orders: Varying the Order

As for finite alphabets, continuity considerations lead to the following extensions of Rényi divergence to orders for which it cannot be defined using the formula in (5).

Definition 3 (Extended Orders). The Rényi divergences of orders 0 and 1 are defined as

$$D_0(P\|Q) = \lim_{\alpha \downarrow 0} D_\alpha(P\|Q),$$

$$D_1(P\|Q) = \lim_{\alpha \uparrow 1} D_\alpha(P\|Q),$$

and of order ∞ as

$$D_\infty(P\|Q) = \lim_{\alpha \uparrow \infty} D_\alpha(P\|Q).$$

Our definition of D_0 follows Csiszár [4]. It differs from Rényi's original definition [1], which uses (5) with $\alpha = 0$ plugged in and is therefore always zero. As illustrated by Section III-F, the present definition is more interesting.

The limits in Definition 3 always exist, because Rényi divergence is nondecreasing in its order:

Theorem 3 (Increasing in the Order). *For $\alpha \in [0, \infty]$ the Rényi divergence $D_\alpha(P\|Q)$ is nondecreasing in α . On $\mathcal{A} = \{\alpha \in [0, \infty] \mid 0 \leq \alpha \leq 1 \text{ or } D_\alpha(P\|Q) < \infty\}$ it is constant if and only if P is the conditional distribution $Q(\cdot \mid A)$ for some event $A \in \mathcal{F}$.*

Proof: Let $\alpha < \beta$ be simple orders. Then for $x \geq 0$ the function $x \mapsto x^{\frac{\alpha-1}{\beta-1}}$ is strictly convex if $\alpha < 1$ and strictly concave if $\alpha > 1$. Therefore by Jensen's inequality

$$\begin{aligned} \frac{1}{\alpha - 1} \ln \int p^\alpha q^{1-\alpha} d\mu &= \frac{1}{\alpha - 1} \ln \int \left(\frac{q}{p}\right)^{(1-\beta)\frac{\alpha-1}{\beta-1}} dP \\ &\leq \frac{1}{\beta - 1} \ln \int \left(\frac{q}{p}\right)^{1-\beta} dP. \end{aligned}$$

On \mathcal{A} , $\int \left(\frac{q}{p}\right)^{1-\beta} dP$ is finite. As a consequence, Jensen's inequality holds with equality if and only if $\left(\frac{q}{p}\right)^{1-\beta}$ is constant P -a.s., which is equivalent to $\frac{q}{p}$ being constant P -a.s., which in turn means that $P = Q(\cdot \mid A)$ for some event A .

From the simple orders, the result extends to the extended orders by the following observations:

$$D_0(P\|Q) = \inf_{0 < \alpha < 1} D_\alpha(P\|Q),$$

$$D_1(P\|Q) = \sup_{0 < \alpha < 1} D_\alpha(P\|Q) \leq \inf_{\alpha > 1} D_\alpha(P\|Q),$$

$$D_\infty(P\|Q) = \sup_{\alpha > 1} D_\alpha(P\|Q).$$

Let us verify that the limits in Definition 3 can be expressed in closed form, just like for finite alphabets. We require the following lemma:

Lemma 1. *For any sequence $\alpha_1, \alpha_2, \dots \in \mathcal{A}$ such that $\alpha_n \rightarrow \beta \in \mathcal{A} \cup \{0, 1\}$*

$$\lim_{n \rightarrow \infty} \int p^{\alpha_n} q^{1-\alpha_n} d\mu = \int \lim_{n \rightarrow \infty} p^{\alpha_n} q^{1-\alpha_n} d\mu. \quad (9)$$

Our proof extends a proof by Shiryaev [25, pp. 366–367].

Proof: We will verify the conditions for the dominated convergence theorem [25], from which (9) follows. First suppose $0 \leq \beta < 1$. Then $0 < \alpha_n < 1$ for all sufficiently large n . In this case $p^{\alpha_n} q^{1-\alpha_n}$, which is never negative, does not exceed $\alpha_n p + (1 - \alpha_n)q \leq p + q$, and the dominated convergence theorem applies because $\int (p + q) d\mu = 2 < \infty$. Secondly, suppose $\beta \geq 1$ and assume without loss of generality that $\alpha_n > 0$. Then there exists a $\gamma \geq \beta$ such that $\gamma \in \mathcal{A} \cup \{1\}$ and $\alpha_n \leq \gamma$ for all sufficiently large n . If $\gamma = 1$, then $\alpha_n < 1$ and we are done by the same argument as above. So suppose $\gamma > 1$. Then convexity of $p^{\alpha_n} q^{1-\alpha_n}$ in α_n implies that for $\alpha_n \leq \gamma$

$$p^{\alpha_n} q^{1-\alpha_n} \leq (1 - \frac{\alpha_n}{\gamma}) p^0 q^1 + \frac{\alpha_n}{\gamma} p^\gamma q^{1-\gamma} \leq q + p^\gamma q^{1-\gamma}.$$

Since $\int q d\mu = 1$, it remains to show that $\int p^\gamma q^{1-\gamma} d\mu < \infty$, which is implied by $\gamma > 1$ and $D_\gamma(P\|Q) < \infty$. ■

The closed-form expression for $\alpha = 0$ follows immediately:

Theorem 4 ($\alpha = 0$).

$$D_0(P\|Q) = -\ln Q(p > 0).$$

Proof of Theorem 4: By Lemma 1 and the fact that $\lim_{\alpha \downarrow 0} p^\alpha q^{1-\alpha} = \mathbf{1}_{\{p>0\}} q$. ■

For $\alpha = 1$, the limit in Definition 3 equals the *Kullback-Leibler divergence* of P from Q , which is defined as

$$D(P\|Q) = \int p \ln \frac{p}{q} d\mu,$$

with the conventions that $0 \ln(\frac{0}{q}) = 0$ and $p \ln(\frac{p}{0}) = \infty$ if $p > 0$. Consequently, $D(P\|Q) = \infty$ if $P \not\ll Q$.

Theorem 5 ($\alpha = 1$).

$$D_1(P\|Q) = D(P\|Q). \quad (10)$$

Moreover, if $D(P\|Q) = \infty$ or there exists a $\beta > 1$ such that $D_\beta(P\|Q) < \infty$, then also

$$\lim_{\alpha \uparrow 1} D_\alpha(P\|Q) = D(P\|Q). \quad (11)$$

For example, by letting $\alpha \uparrow 1$ in (6) or by direct computation, it can be derived [19] that the Kullback-Leibler divergence between two normal distributions with positive variance is

$$\begin{aligned} D_1(\mathcal{N}(\mu_0, \sigma_0^2) \| \mathcal{N}(\mu_1, \sigma_1^2)) \\ = \frac{1}{2} \left(\frac{(\mu_1 - \mu_0)^2}{\sigma_1^2} + \ln \frac{\sigma_1^2}{\sigma_0^2} + \frac{\sigma_0^2}{\sigma_1^2} - 1 \right). \end{aligned}$$

It is possible that $D_\alpha(P\|Q) = \infty$ for all $\alpha > 1$, but $D(P\|Q) < \infty$, such that (11) does not hold. This situation occurs, for example, if P is doubly exponential on $\mathcal{X} = \mathbb{R}$ with density $p(x) = e^{-2|x|}$ and Q is standard normal with density $q(x) = e^{-x^2/2}/\sqrt{\tau}$, where $\tau = 2\pi$. (Liese and Vajda [24] have previously used these distributions in a similar example.) In this case there is no way to make Rényi divergence continuous in α at $\alpha = 1$, and we opt to define D_1 as the limit from below, such that it always equals the Kullback-Leibler divergence.

The proof of Theorem 5 requires an intermediate lemma:

Lemma 2. For any $x > 1/2$

$$(x-1) \left(1 + \frac{1-x}{2} \right) \leq \ln x \leq x-1.$$

Proof: By Taylor's theorem with Cauchy's remainder term we have for any positive x that

$$\begin{aligned} \ln x &= x-1 - \frac{(x-\xi)(x-1)}{2\xi^2} \\ &= (x-1) \left(1 + \frac{\xi-x}{2\xi^2} \right) \end{aligned}$$

for some ξ between x and 1. As $\frac{\xi-x}{2\xi^2}$ is increasing in ξ for $x > \frac{1}{2}$, the lemma follows. ■

Proof of Theorem 5: Suppose $P \not\ll Q$. Then $D(P\|Q) = \infty = D_\beta(P\|Q)$ for all $\beta > 1$, so (11) holds. And (10) follows by

$$\begin{aligned} \lim_{\alpha \uparrow 1} \frac{1}{\alpha-1} \ln \int p^\alpha q^{1-\alpha} d\mu &\geq \lim_{\alpha \uparrow 1} \frac{1}{\alpha-1} \ln \int (\mathbf{1}_{\{q>0\}} p)^\alpha d\mu \\ &\geq \lim_{\alpha \uparrow 1} \frac{\alpha}{\alpha-1} \ln P(q > 0) = \infty = D(P\|Q), \end{aligned}$$

where the second inequality is Jensen's. Alternatively, suppose $P \ll Q$ and let $x_\alpha = \int p^\alpha q^{1-\alpha} d\mu$. Then $\lim_{\alpha \uparrow 1} x_\alpha = 1$ by Lemma 1. Therefore Lemma 2 implies that

$$\begin{aligned} \lim_{\alpha \uparrow 1} D_\alpha(P\|Q) &= \lim_{\alpha \uparrow 1} \frac{1}{\alpha-1} \ln x_\alpha \\ &= \lim_{\alpha \uparrow 1} \frac{x_\alpha - 1}{\alpha-1} = \lim_{\alpha \uparrow 1} \int_{p,q>0} \frac{p - p^\alpha q^{1-\alpha}}{1-\alpha} d\mu, \quad (12) \end{aligned}$$

where the restriction of the domain of integration is allowed because $q = 0$ implies $p = 0$ (μ -a.s.) by $P \ll Q$. Convexity of $p^\alpha q^{1-\alpha}$ in α implies that its derivative, $p^\alpha q^{1-\alpha} \ln \frac{p}{q}$, is nondecreasing and therefore for $p, q > 0$

$$\frac{p - p^\alpha q^{1-\alpha}}{1-\alpha} = \frac{1}{1-\alpha} \int_\alpha^1 p^z q^{1-z} \ln \frac{p}{q} dz$$

is nondecreasing in α , and $\frac{p - p^\alpha q^{1-\alpha}}{1-\alpha} \geq \frac{p - p^0 q^{1-0}}{1-0} = p - q$. As $\int_{p,q>0} (p - q) d\mu > -\infty$, it follows by the monotone convergence theorem that

$$\begin{aligned} \lim_{\alpha \uparrow 1} \int_{p,q>0} \frac{p - p^\alpha q^{1-\alpha}}{1-\alpha} d\mu &= \int_{p,q>0} \lim_{\alpha \uparrow 1} \frac{p - p^\alpha q^{1-\alpha}}{1-\alpha} d\mu \\ &= \int_{p,q>0} p \ln \frac{p}{q} d\mu = D(P\|Q), \end{aligned}$$

which together with (12) proves (10). If $D(P\|Q) = \infty$, then $D_\beta(P\|Q) \geq D(P\|Q) = \infty$ for all $\beta > 1$ and (11) holds. It remains to prove (11) if there exists a $\beta > 1$ such that $D_\beta(P\|Q) < \infty$. In this case, arguments similar to the ones above imply that

$$\lim_{\alpha \downarrow 1} D_\alpha(P\|Q) = \lim_{\alpha \downarrow 1} \int_{p,q>0} \frac{p^\alpha q^{1-\alpha} - p}{\alpha-1} d\mu \quad (13)$$

and $\frac{p^\alpha q^{1-\alpha} - p}{\alpha-1}$ is nondecreasing in α . Therefore $\frac{p^\alpha q^{1-\alpha} - p}{\alpha-1} \leq \frac{p^\beta q^{1-\beta} - p}{\beta-1} \leq \frac{p^\beta q^{1-\beta}}{\beta-1}$ and, as $\int_{p,q>0} \frac{p^\beta q^{1-\beta}}{\beta-1} d\mu < \infty$ is implied

by $D_\beta(P\|Q) < \infty$, it follows by the monotone convergence theorem that

$$\begin{aligned} \lim_{\alpha \downarrow 1} \int_{p,q>0} \frac{p^\alpha q^{1-\alpha} - p}{\alpha - 1} d\mu &= \int_{p,q>0} \lim_{\alpha \downarrow 1} \frac{p^\alpha q^{1-\alpha} - p}{\alpha - 1} d\mu \\ &= \int_{p,q>0} p \ln \frac{p}{q} d\mu = D(P\|Q), \end{aligned}$$

which together with (13) completes the proof. ■

For any random variable X , the *essential supremum* of X with respect to P is $\text{ess sup}_P X = \inf\{c \mid P(X > c) = 0\}$.

Theorem 6 ($\alpha = \infty$).

$$D_\infty(P\|Q) = \ln \sup_{A \in \mathcal{F}} \frac{P(A)}{Q(A)} = \ln \left(\text{ess sup}_P \frac{p}{q} \right),$$

with the conventions that $0/0 = 0$ and $x/0 = \infty$ if $x > 0$.

If the sample space \mathcal{X} is countable, then with the notational conventions of this theorem the essential supremum reduces to an ordinary supremum, and we have $D_\infty(P\|Q) = \ln \sup_x \frac{P(x)}{Q(x)}$.

Proof: If \mathcal{X} contains a finite number of elements n , then

$$\begin{aligned} D_\infty(P\|Q) &= \lim_{\alpha \uparrow \infty} \frac{1}{\alpha - 1} \ln \sum_{i=1}^n p_i^\alpha q_i^{1-\alpha} \\ &= \ln \max_i \frac{p_i}{q_i} = \ln \max_{A \subseteq \mathcal{X}} \frac{P(A)}{Q(A)}. \end{aligned}$$

This extends to arbitrary measurable spaces $(\mathcal{X}, \mathcal{F})$ by Theorem 2:

$$\begin{aligned} D_\infty(P\|Q) &= \sup_{\alpha < \infty} \sup_{\mathcal{P}} D_\alpha(P_{|\mathcal{P}}\|Q_{|\mathcal{P}}) = \sup_{\mathcal{P}} \sup_{\alpha < \infty} D_\alpha(P_{|\mathcal{P}}\|Q_{|\mathcal{P}}) \text{ Jensen's inequality} \\ &= \sup_{\mathcal{P}} \ln \max_{A \in \mathcal{P}} \frac{P(A)}{Q(A)} = \ln \sup_{A \in \mathcal{F}} \frac{P(A)}{Q(A)}, \end{aligned}$$

where \mathcal{P} ranges over all finite partitions in \mathcal{F} .

Now if $P \not\ll Q$, then there exists an event $B \in \mathcal{F}$ such that $P(B) > 0$ but $Q(B) = 0$, and

$$P\left(\frac{p}{q} = \infty\right) = P(q = 0) \geq P(B) > 0$$

implies that $\text{ess sup } p/q = \infty = \sup_A P(A)/Q(A)$. Alternatively, suppose that $P \ll Q$. Then

$$P(A) = \int_{A \cap \{q>0\}} p d\mu \leq \int_{A \cap \{q>0\}} \text{ess sup} \frac{p}{q} \cdot q d\mu = \text{ess sup} \frac{p}{q} \cdot Q(A)$$

for all $A \in \mathcal{F}$ and it follows that

$$\sup_{A \in \mathcal{F}} \frac{P(A)}{Q(A)} \leq \text{ess sup} \frac{p}{q}. \quad (14)$$

Let $a < \text{ess sup } p/q$ be arbitrary. Then there exists a set $A \in \mathcal{F}$ with $P(A) > 0$ such that $\frac{p}{q} \geq a$ on A and therefore

$$P(A) = \int_A p d\mu \geq \int_A a \cdot q d\mu = a \cdot Q(A).$$

Thus $\sup_{A \in \mathcal{F}} \frac{P(A)}{Q(A)} \geq a$ for any $a < \text{ess sup } \frac{p}{q}$, which implies that

$$\sup_{A \in \mathcal{F}} \frac{P(A)}{Q(A)} \geq \text{ess sup} \frac{p}{q}.$$

In combination with (14) this completes the proof. ■

Taken together, the previous results imply that Rényi divergence is a continuous function of its order α (under suitable conditions):

Theorem 7 (Continuity in the Order). *The Rényi divergence $D_\alpha(P\|Q)$ is continuous in α on $A = \{\alpha \in [0, \infty] \mid 0 \leq \alpha \leq 1 \text{ or } D_\alpha(P\|Q) < \infty\}$.*

Proof: Continuity at any simple order β follows by Lemma 1. It extends to the extended orders 0 and ∞ by the definition of Rényi divergence at these orders. And it extends to $\alpha = 1$ by Theorem 5. ■

III. FIXED NONNEGATIVE ORDERS

In this section we fix the order α and study properties of Rényi divergence as P and Q are varied. First we prove nonnegativity and extend the data processing inequality and the relation to a supremum over finite partitions to extended orders. Then we consider various convexity and continuity properties.

A. Positivity, Data Processing and Finite Partitions

Theorem 8 (Positivity). *For any order $\alpha \in [0, \infty]$*

$$D_\alpha(P\|Q) \geq 0.$$

For $\alpha > 0$, $D_\alpha(P\|Q) = 0$ if and only if $P = Q$. For $\alpha = 0$, $D_\alpha(P\|Q) = 0$ if and only if $Q \ll P$.

Proof: Suppose first that α is a simple order. Then by

$$\begin{aligned} \frac{1}{\alpha - 1} \ln \int p^\alpha q^{1-\alpha} d\mu &= \frac{1}{\alpha - 1} \ln \int \left(\frac{q}{p}\right)^{1-\alpha} dP \\ &\geq \frac{1 - \alpha}{\alpha - 1} \ln \int \frac{q}{p} dP \geq 0. \end{aligned}$$

Equality holds if and only if q/p is constant P -a.s. (first inequality) and $Q \ll P$ (second inequality), which together is equivalent to $P = Q$.

The result extends to $\alpha \in \{1, \infty\}$ by $D_\alpha(P\|Q) = \sup_{\beta < \alpha} D_\beta(P\|Q)$. For $\alpha = 0$ it can be verified directly that $-\ln Q(p > 0) \geq 0$, with equality if and only if $Q \ll P$. ■

Theorem 9 (Data Processing Inequality). *For any order $\alpha \in [0, \infty]$ and any σ -subalgebra $\mathcal{G} \subseteq \mathcal{F}$*

$$D_\alpha(P_{|\mathcal{G}}\|Q_{|\mathcal{G}}) \leq D_\alpha(P\|Q). \quad (15)$$

Proof: By Theorem 1, (15) holds for the simple orders. Let β be any extended order and let $\alpha_n \rightarrow \beta$ be an arbitrary sequence of simple orders that converges to β , from above if $\beta = 0$ and from below if $\beta \in \{1, \infty\}$. Then

$$\begin{aligned} D_\beta(P_{|\mathcal{G}}\|Q_{|\mathcal{G}}) &= \lim_{n \rightarrow \infty} D_{\alpha_n}(P_{|\mathcal{G}}\|Q_{|\mathcal{G}}) \\ &\leq \lim_{n \rightarrow \infty} D_{\alpha_n}(P\|Q) = D_\beta(P\|Q). \end{aligned}$$

Theorem 10. *For any $\alpha \in (0, \infty]$*

$$D_\alpha(P\|Q) = \sup_{\mathcal{P}} D_\alpha(P_{|\mathcal{P}}\|Q_{|\mathcal{P}}),$$

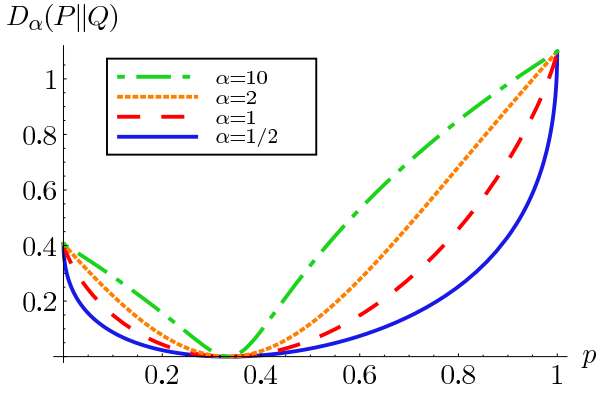


Fig. 2. Rényi divergence as a function of $P = (p, 1-p)$ for $Q = (1/3, 2/3)$

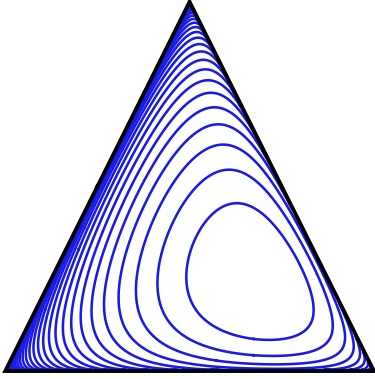


Fig. 3. Level curves of $D_{1/2}(P||Q)$ for fixed Q as P ranges over the simplex of distributions on a three-element set

where the supremum is over all finite partitions $\mathcal{P} \subseteq \mathcal{F}$.

Proof: For simple orders α , the result holds by Theorem 2. This extends to $\alpha \in \{1, \infty\}$ by monotonicity and left-continuity in α :

$$\begin{aligned} D_\alpha(P||Q) &= \sup_{\beta < \alpha} D_\beta(P||Q) = \sup_{\beta < \alpha} \sup_{\mathcal{P}} D_\beta(P_{|\mathcal{P}}||Q_{|\mathcal{P}}) \\ &= \sup_{\mathcal{P}} \sup_{\beta < \alpha} D_\beta(P_{|\mathcal{P}}||Q_{|\mathcal{P}}) = \sup_{\mathcal{P}} D_\alpha(P_{|\mathcal{P}}||Q_{|\mathcal{P}}). \end{aligned}$$

B. Convexity

Consider Figures 2 and 3. They show $D_\alpha(P||Q)$ as a function of P for sample spaces containing two or three elements. These figures suggest that Rényi divergence is convex in its first argument for small α , but not for large α . This is in agreement with the well-known fact that it is jointly convex in the pair (P, Q) for $\alpha = 1$. It turns out that joint convexity extends to $\alpha < 1$, but not to $\alpha > 1$, as noted by Csiszár [4]. Our proof generalises the proof for $\alpha = 1$ by Cover and Thomas [27].

Theorem 11. *For any order $\alpha \in [0, 1]$ Rényi divergence is jointly convex in its arguments. That is, for any two pairs of probability distributions (P_0, Q_0) and (P_1, Q_1) , and any*

$$0 < \lambda < 1$$

$$\begin{aligned} D_\alpha((1-\lambda)P_0 + \lambda P_1 || (1-\lambda)Q_0 + \lambda Q_1) \\ \leq (1-\lambda)D_\alpha(P_0||Q_0) + \lambda D_\alpha(P_1||Q_1). \end{aligned} \quad (16)$$

Equality holds if and only if

$$\begin{aligned} \alpha = 0: & D_0(P_0||Q_0) = D_0(P_1||Q_1), \\ & p_0 = 0 \Rightarrow p_1 = 0 \text{ (} Q_0\text{-a.s.) and} \\ & p_1 = 0 \Rightarrow p_0 = 0 \text{ (} Q_1\text{-a.s.)}; \\ 0 < \alpha < 1: & D_\alpha(P_0||Q_0) = D_\alpha(P_1||Q_1) \text{ and} \\ & p_0 q_1 = p_1 q_0 \text{ (}\mu\text{-a.s.)}; \\ \alpha = 1: & p_0 q_1 = p_1 q_0 \text{ (}\mu\text{-a.s.)} \end{aligned}$$

Proof: Suppose first that $\alpha = 0$, and let $P_\lambda = (1-\lambda)P_0 + \lambda P_1$ and $Q_\lambda = (1-\lambda)Q_0 + \lambda Q_1$. Then

$$\begin{aligned} (1-\lambda) \ln Q_0(p_0 > 0) + \lambda \ln Q_1(p_1 > 0) \\ \leq \ln((1-\lambda)Q_0(p_0 > 0) + \lambda Q_1(p_1 > 0)) \\ \leq \ln Q_\lambda(p_0 > 0 \text{ or } p_1 > 0) = \ln Q_\lambda(p_\lambda > 0). \end{aligned}$$

Equality holds if and only if, for the first inequality, $Q_0(p_0 > 0) = Q_1(p_1 > 0)$ and, for the second inequality, $p_1 > 0 \Rightarrow p_0 > 0$ (Q_0 -a.s.) and $p_0 > 0 \Rightarrow p_1 > 0$ (Q_1 -a.s.) These conditions are equivalent to the equality conditions of the theorem.

Alternatively, suppose $\alpha > 0$. We will show that pointwise

$$\begin{aligned} (1-\lambda)p_0^\alpha q_0^{1-\alpha} + \lambda p_1^\alpha q_1^{1-\alpha} &\leq p_\lambda^\alpha q_\lambda^{1-\alpha} \quad (0 < \alpha < 1); \\ (1-\lambda)p_0 \ln \frac{p_0}{q_0} + \lambda p_1 \ln \frac{p_1}{q_1} &\geq p_\lambda \ln \frac{p_\lambda}{q_\lambda} \quad (\alpha = 1), \end{aligned} \quad (17)$$

where $p_\lambda = (1-\lambda)p_0 + \lambda p_1$ and $q_\lambda = (1-\lambda)q_0 + \lambda q_1$. For $\alpha = 1$, (16) then follows directly; for $0 < \alpha < 1$, (16) follows from (17) by Jensen's inequality:

$$\begin{aligned} (1-\lambda) \ln \int p_0^\alpha q_0^{1-\alpha} d\mu + \lambda \ln \int p_1^\alpha q_1^{1-\alpha} d\mu \\ \leq \ln \left((1-\lambda) \int p_0^\alpha q_0^{1-\alpha} d\mu + \lambda \int p_1^\alpha q_1^{1-\alpha} d\mu \right). \end{aligned} \quad (18)$$

If one of p_0, p_1, q_0 and q_1 is zero, then (17) can be verified directly. So assume that they are all positive. Then for $0 < \alpha < 1$ let $f(x) = -x^\alpha$ and for $\alpha = 1$ let $f(x) = x \ln x$, such that (17) can be written as

$$\frac{(1-\lambda)q_0}{q_\lambda} f\left(\frac{p_0}{q_0}\right) + \frac{\lambda q_1}{q_\lambda} f\left(\frac{p_1}{q_1}\right) \geq f\left(\frac{p_\lambda}{q_\lambda}\right).$$

(17) is established by recognising this as an application of Jensen's inequality to the strictly convex function f . Regardless of whether any of p_0, p_1, q_0 and q_1 is zero, equality holds in (17) if and only if $p_0 q_1 = p_1 q_0$. Equality holds in (18) if and only if $\int p_0^\alpha q_0^{1-\alpha} d\mu = \int p_1^\alpha q_1^{1-\alpha} d\mu$, which is equivalent to $D_\alpha(P_0||Q_0) = D_\alpha(P_1||Q_1)$. ■

Joint convexity in P and Q breaks down for $\alpha > 1$ (see Appendix B), but some partial convexity properties can still be salvaged. First, convexity in the second argument does hold for all α [4]:

Theorem 12. For any order $\alpha \in [0, \infty]$ Rényi divergence is convex in its second argument. That is, for any probability distributions P , Q_0 and Q_1

$$D_\alpha(P \parallel (1-\lambda)Q_0 + \lambda Q_1) \leq (1-\lambda)D_\alpha(P \parallel Q_0) + \lambda D_\alpha(P \parallel Q_1) \quad (19)$$

for any $0 < \lambda < 1$. For finite α , equality holds if and only if

$$\alpha = 0: D_0(P \parallel Q_0) = D_0(P \parallel Q_1);$$

$$0 < \alpha < \infty: q_0 = q_1 \text{ (P-a.s.)}$$

Proof: For $\alpha \in [0, 1]$ this follows from the previous theorem. (For $P_0 = P_1$ the equality conditions reduce to the ones given here.) For $\alpha \in (1, \infty)$, let $Q_\lambda = (1-\lambda)Q_0 + \lambda Q_1$ and define $f(x, Q_\lambda) = (p(x)/q_\lambda(x))^{\alpha-1}$. It is sufficient to show that

$$\begin{aligned} & \ln \mathbf{E}_{X \sim P}[f(X, Q_\lambda)] \\ & \leq (1-\lambda) \ln \mathbf{E}_{X \sim P}[f(X, Q_0)] + \lambda \ln \mathbf{E}_{X \sim P}[f(X, Q_1)]. \end{aligned}$$

Noting that, for every $x \in \mathcal{X}$, $f(x, Q)$ is log-convex in Q , this is a consequence of the general fact that an expectation over log-convex functions is itself log-convex, which can be shown using Hölder's inequality:

$$\begin{aligned} \mathbf{E}_P[f(X, Q_\lambda)] & \leq \mathbf{E}_P[f(X, Q_0)^{1-\lambda} f(X, Q_1)^\lambda] \\ & \leq \mathbf{E}_P[f(X, Q_0)]^{1-\lambda} \mathbf{E}_P[f(X, Q_1)]^\lambda. \end{aligned}$$

Taking logarithms completes the proof of (19). Equality holds in the first inequality if and only if $q_0 = q_1$ (P-a.s.), which is also sufficient for equality in the second inequality. Finally, (19) extends to $\alpha = \infty$ by letting α tend to ∞ . ■

And secondly, Rényi divergence is jointly *quasi-convex* in both arguments for all α :

Theorem 13. For any order $\alpha \in [0, \infty]$ Rényi divergence is jointly quasi-convex in its arguments. That is, for any two pairs of probability distributions (P_0, Q_0) and (P_1, Q_1) , and any $\lambda \in (0, 1)$

$$\begin{aligned} D_\alpha((1-\lambda)P_0 + \lambda P_1 \parallel (1-\lambda)Q_0 + \lambda Q_1) \\ \leq \max\{D_\alpha(P_0 \parallel Q_0), D_\alpha(P_1 \parallel Q_1)\}. \end{aligned} \quad (20)$$

Proof: For $\alpha \in [0, 1]$, quasi-convexity is implied by convexity. For $\alpha \in (1, \infty)$, strict monotonicity of $x \mapsto \frac{1}{\alpha-1} \ln x$ implies that quasi-convexity is equivalent to quasi-convexity of the Hellinger integral $\int p^\alpha q^{1-\alpha} d\mu$. Since quasi-convexity is implied by ordinary convexity, it is sufficient to establish that the Hellinger integral is jointly convex in P and Q . Let $p_\lambda = (1-\lambda)p_0 + \lambda p_1$ and $q_\lambda = (1-\lambda)q_0 + \lambda q_1$. Then joint convexity of the Hellinger integral is implied by the pointwise inequality

$$(1-\lambda)p_0^\alpha q_0^{1-\alpha} + \lambda p_1^\alpha q_1^{1-\alpha} \geq p_\lambda^\alpha q_\lambda^{1-\alpha},$$

which holds by essentially the same argument as for (17) in the proof of Theorem 11, with the convex function $f(x) = x^\alpha$.

Finally, the case $\alpha = \infty$ follows by letting α tend to ∞ :

$$\begin{aligned} D_\infty((1-\lambda)P_0 + \lambda P_1 \parallel (1-\lambda)Q_0 + \lambda Q_1) \\ = \sup_{\alpha < \infty} D_\alpha((1-\lambda)P_0 + \lambda P_1 \parallel (1-\lambda)Q_0 + \lambda Q_1) \\ \leq \sup_{\alpha < \infty} \max\{D_\alpha(P_0 \parallel Q_0), D_\alpha(P_1 \parallel Q_1)\} \\ = \max\{\sup_{\alpha < \infty} D_\alpha(P_0 \parallel Q_0), \sup_{\alpha < \infty} D_\alpha(P_1 \parallel Q_1)\} \\ = \max\{D_\infty(P_0 \parallel Q_0), D_\infty(P_1 \parallel Q_1)\}. \end{aligned}$$

■

C. Continuity

In this section we study continuity properties of the Rényi divergence $D_\alpha(P \parallel Q)$ of different orders in the pair of probability distributions (P, Q) . It turns out that continuity depends on the order α and the topology on the set of all probability distributions.

If the set of probability distributions on $(\mathcal{X}, \mathcal{F})$ is equipped with the topology of setwise convergence (τ -topology), then convergence of a sequence of probability distributions P_1, P_2, \dots to a probability distribution Q means that $P_n(A) \rightarrow Q(A)$ for any $A \in \mathcal{F}$. Alternatively, one might consider the topology defined by the *total variation distance*

$$V(P, Q) = \int |p - q| d\mu = 2 \sup_{A \in \mathcal{F}} |P(A) - Q(A)|,$$

in which $P_n \rightarrow Q$ means that $V(P_n, Q) \rightarrow 0$. The total variation topology is stronger than the topology of setwise convergence in the sense that convergence in total variation distance implies convergence on any $A \in \mathcal{F}$. The two topologies coincide if the sample space \mathcal{X} is countable.

In general, Rényi divergence is lower semi-continuous for positive orders:

Theorem 14. For any order $\alpha \in (0, \infty]$, $D_\alpha(P \parallel Q)$ is a lower semi-continuous function of the pair (P, Q) in the topology of setwise convergence.

Proof: Suppose $\mathcal{X} = \{a_1, \dots, a_k\}$ is finite. Then for any simple order α

$$D_\alpha(P \parallel Q) = \frac{1}{\alpha-1} \ln \sum_{i=1}^k p_i^\alpha q_i^{1-\alpha},$$

where $p_i = P(a_i)$ and $q_i = Q(a_i)$. If $0 < \alpha < 1$, then $p_i^\alpha q_i^{1-\alpha}$ is continuous in (P, Q) . For $1 < \alpha < \infty$, it is only discontinuous at $p_i = q_i = 0$, but there $p_i^\alpha q_i^{1-\alpha} = 0 = \min_{(P, Q)} p_i^\alpha q_i^{1-\alpha}$, so then $p_i^\alpha q_i^{1-\alpha}$ is still lower semi-continuous. These properties carry over to $\sum_{i=1}^k p_i^\alpha q_i^{1-\alpha}$ and thus $D_\alpha(P \parallel Q)$ is continuous for $0 < \alpha < 1$ and lower semi-continuous for $\alpha > 1$. A supremum over (lower semi-)continuous functions is itself lower semi-continuous. Therefore, for simple orders α , Theorem 2 implies that $D_\alpha(P \parallel Q)$ is lower semi-continuous for arbitrary \mathcal{X} . This property extends to the extended orders 1 and ∞ by $D_\beta(P \parallel Q) = \sup_{\alpha < \beta} D_\alpha(P \parallel Q)$ for $\beta \in \{1, \infty\}$. ■

Moreover, if $\alpha \in (0, 1)$ and the stronger of the two topologies is assumed, then Theorem 16 below shows that Rényi divergence is continuous.

First we prove that the topologies induced by Rényi divergences of orders $\alpha \in (0, 1)$ are all equivalent:

Theorem 15. *For any $0 < \alpha \leq \beta < 1$*

$$\frac{\alpha}{\beta} \frac{1-\beta}{1-\alpha} D_\beta(P\|Q) \leq D_\alpha(P\|Q) \leq D_\beta(P\|Q).$$

This follows from the following symmetry-like property, which may be verified directly.

Proposition 2 (Skew Symmetry). *For any $0 < \alpha < 1$*

$$D_\alpha(P\|Q) = \frac{\alpha}{1-\alpha} D_{1-\alpha}(Q\|P).$$

Note that, in particular, Rényi divergence is symmetric for $\alpha = \frac{1}{2}$, but that skew symmetry does not hold for $\alpha = 0$ and $\alpha = 1$.

Proof of Theorem 15: We have already established the second inequality in Theorem 3, so it remains to prove the first one. Skew symmetry implies that

$$\begin{aligned} \frac{1-\alpha}{\alpha} D_\alpha(P\|Q) &= D_{1-\alpha}(Q\|P) \\ &\geq D_{1-\beta}(Q\|P) = \frac{1-\beta}{\beta} D_\beta(P\|Q), \end{aligned}$$

from which the result follows. ■

By (2), these results show that, for $\alpha \in (0, 1)$, $D_\alpha(P_n\|Q) \rightarrow 0$ is equivalent to convergence of P_n to Q in Hellinger distance, which is equivalent to convergence of P_n to Q in total variation [25, p. 364]. Next we shall prove a stronger result on the relation between Rényi divergence and total variation.

Theorem 16. *For $\alpha \in (0, 1)$, the Rényi divergence $D_\alpha(P\|Q)$ is a continuous function of (P, Q) in the total variation topology.*

Lemma 3. *Let $0 < \alpha < 1$. Then for all $x, y \geq 0$ and $\varepsilon > 0$*

$$|x^\alpha - y^\alpha| \leq \varepsilon^\alpha + \varepsilon^{\alpha-1} |x - y|.$$

Proof: If $x, y \leq \varepsilon$ or $x = y$ the inequality $|x^\alpha - y^\alpha| \leq \varepsilon^\alpha$ is obvious. So assume that $x > y$ and $x \geq \varepsilon$. Then

$$\frac{|x^\alpha - y^\alpha|}{|x - y|} \leq \frac{|x^\alpha - 0^\alpha|}{|x - 0|} = x^{\alpha-1} \leq \varepsilon^{\alpha-1}.$$

Proof of Theorem 16: First note that Rényi divergence is a function of the power divergence $d_\alpha(P, Q) = \int \left(1 - \left(\frac{dP}{dQ}\right)^\alpha\right) dQ$:

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1} \ln(1 - d_\alpha(P, Q)).$$

Since $x \mapsto \frac{1}{\alpha-1} \ln(1-x)$ is continuous, it is sufficient to prove that $d_\alpha(P, Q)$ is a continuous function of (P, Q) . For any $\varepsilon > 0$ and distributions P_1, P_2 and Q , Lemma 3 implies

that

$$\begin{aligned} |d_\alpha(P_1, Q) - d_\alpha(P_2, Q)| &\leq \int \left| \left(\frac{dP_1}{dQ}\right)^\alpha - \left(\frac{dP_2}{dQ}\right)^\alpha \right| dQ \\ &\leq \int \left(\varepsilon^\alpha + \varepsilon^{\alpha-1} \left| \frac{dP_1}{dQ} - \frac{dP_2}{dQ} \right| \right) dQ \\ &= \varepsilon^\alpha + \varepsilon^{\alpha-1} \int \left| \frac{dP_1}{dQ} - \frac{dP_2}{dQ} \right| dQ \\ &= \varepsilon^\alpha + \varepsilon^{\alpha-1} V(P_1, P_2). \end{aligned}$$

As $d_\alpha(P, Q) = d_{1-\alpha}(Q, P)$, it also follows that $|d_\alpha(P, Q_1) - d_\alpha(P, Q_2)| \leq \varepsilon^{1-\alpha} + \varepsilon^{-\alpha} V(Q_1, Q_2)$ for any Q_1, Q_2 and P . Therefore

$$\begin{aligned} |d_\alpha(P_1, Q_1) - d_\alpha(P_2, Q_2)| &\leq |d_\alpha(P_1, Q_1) - d_\alpha(P_2, Q_1)| \\ &\quad + |d_\alpha(P_2, Q_1) - d_\alpha(P_2, Q_2)| \\ &\leq \varepsilon^\alpha + \varepsilon^{\alpha-1} V(P_1, P_2) + \varepsilon^{1-\alpha} + \varepsilon^{-\alpha} V(Q_1, Q_2), \end{aligned}$$

from which the theorem follows. ■

A partial extension to $\alpha = 0$ follows:

Corollary 1. *The Rényi divergence $D_0(P\|Q)$ is an upper semi-continuous function of (P, Q) in the total variation topology.*

Proof: This follows from Theorem 16 because $D_0(P\|Q)$ is the infimum of the continuous functions $(P, Q) \mapsto D_\alpha(P\|Q)$ for $\alpha \in (0, 1)$. ■

Finally, if we consider continuity in Q only, we obtain:

Theorem 17. *Suppose \mathcal{X} is finite, and let $\alpha \in [0, \infty]$. Then for any P the Rényi divergence $D_\alpha(P\|Q)$ is continuous in Q .*

Proof: Directly from the closed-form expressions for Rényi divergence. ■

D. Limits of σ -Algebras

As shown by Theorem 2, there exists a sequence of finite partitions $\mathcal{P}_1, \mathcal{P}_2, \dots$ such that

$$D_\alpha(P|_{\mathcal{P}_n}\|Q|_{\mathcal{P}_n}) \uparrow D_\alpha(P\|Q). \quad (21)$$

Theorem 18 below elaborates on this result. It implies that (21) holds for any increasing sequence of partitions $\mathcal{P}_1 \subseteq \mathcal{P}_2 \subseteq \dots$ that generate σ -algebras converging to \mathcal{F} , in the sense that $\mathcal{F} = \sigma(\bigcup_{n=1}^\infty \mathcal{P}_n)$. A corresponding result holds for infinite sequences of increasingly coarse partitions, which is shown by Theorem 19.

Theorem 18 (Increasing). *Let $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots \subseteq \mathcal{F}$ be a non-decreasing family of σ -algebras, and let $\mathcal{F}_\infty = \sigma(\bigcup_{n=1}^\infty \mathcal{F}_n)$ be the smallest σ -algebra containing them. Then for any order $\alpha \in (0, \infty]$*

$$\lim_{n \rightarrow \infty} D_\alpha(P|_{\mathcal{F}_n}\|Q|_{\mathcal{F}_n}) = D_\alpha(P|_{\mathcal{F}_\infty}\|Q|_{\mathcal{F}_\infty}). \quad (22)$$

For $\alpha = 0$, (22) does not hold. A counterexample is given after Example 2 below.

Lemma 4. *Let $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots \subseteq \mathcal{F}$ be a nondecreasing family of σ -algebras, and let P and μ be probability distributions on*

$(\mathcal{X}, \mathcal{F})$ such that $P \ll \mu$. Let p be the density of P with respect to μ . Then the family of random variables $\{X_n\}_{n \geq 1}$ with members $X_n = \mathbf{E}[p | \mathcal{F}_n]$ is uniformly integrable (with respect to μ).

The proof of this lemma is a special case of part of the proof of Lévy's theorem in Shiryaev's textbook [25]. We repeat it here for completeness.

Proof: For any constants $b, c > 0$

$$\begin{aligned} \int_{X_n > b} X_n d\mu &= \int_{X_n > b} p d\mu \\ &\leq \int_{X_n > b, p \leq c} p d\mu + \int_{X_n > b, p > c} p d\mu \\ &\leq c \cdot \mu(X_n > b) + \int_{p > c} p d\mu \\ &\stackrel{(*)}{\leq} \frac{c}{b} \mathbf{E}[X_n] + \int_{p > c} p d\mu = \frac{c}{b} + \int_{p > c} p d\mu, \end{aligned}$$

in which the inequality marked by $(*)$ is Markov's. Consequently

$$\begin{aligned} \limsup_{b \rightarrow \infty} \int_{X_n > b} |X_n| d\mu &= \lim_{c \rightarrow \infty} \limsup_{b \rightarrow \infty} \int_{X_n > b} |X_n| d\mu \\ &\leq \lim_{c \rightarrow \infty} \lim_{b \rightarrow \infty} \frac{c}{b} + \lim_{c \rightarrow \infty} \int_{p > c} p d\mu = 0, \end{aligned}$$

which proves the lemma. \blacksquare

Proof of Theorem 18: As by the data processing inequality $D_\alpha(P_{|\mathcal{F}_n} \| Q_{|\mathcal{F}_n}) \leq D_\alpha(P \| Q)$ for all n , we only need to show that $\lim_{n \rightarrow \infty} D_\alpha(P_{|\mathcal{F}_n} \| Q_{|\mathcal{F}_n}) \geq D_\alpha(P_{|\mathcal{F}_\infty} \| Q_{|\mathcal{F}_\infty})$. To this end, assume without loss of generality that $\mathcal{F} = \mathcal{F}_\infty$ and that μ is a probability distribution (i.e. $\mu = (P + Q)/2$). Let $X_n = \mathbf{E}[p | \mathcal{F}_n]$ and $Y_n = \mathbf{E}[q | \mathcal{F}_n]$, and define the distributions \tilde{P}_n and \tilde{Q}_n on $(\mathcal{X}, \mathcal{F})$ by

$$\tilde{P}_n(A) = \int_A X_n d\mu, \quad \tilde{Q}_n(A) = \int_A Y_n d\mu \quad (A \in \mathcal{F}),$$

such that, by the Radon-Nikodým theorem and Proposition 1, $\frac{d\tilde{P}_n}{d\mu} = X_n = \frac{dP_{|\mathcal{F}_n}}{d\mu_{|\mathcal{F}_n}}$ and $\frac{d\tilde{Q}_n}{d\mu} = Y_n = \frac{dQ_{|\mathcal{F}_n}}{d\mu_{|\mathcal{F}_n}}$ (μ -a.s.) It follows that

$$D_\alpha(\tilde{P}_n \| \tilde{Q}_n) = D_\alpha(P_{|\mathcal{F}_n} \| Q_{|\mathcal{F}_n})$$

for $0 < \alpha < \infty$ and therefore by continuity also for $\alpha = \infty$. We will proceed to show that $(\tilde{P}_n, \tilde{Q}_n) \rightarrow (P, Q)$ in the topology of setwise convergence. By lower semi-continuity of Rényi divergence this implies that $\lim_{n \rightarrow \infty} D_\alpha(\tilde{P}_n \| \tilde{Q}_n) \geq D_\alpha(P \| Q)$, from which the theorem follows. By Lévy's theorem [25], $\lim_{n \rightarrow \infty} X_n = p$ (μ -a.s.) Hence uniform integrability of the family $\{X_n\}$ (by Lemma 4) implies that for any $A \in \mathcal{F}$

$$\lim_{n \rightarrow \infty} \tilde{P}_n(A) = \lim_{n \rightarrow \infty} \int_A X_n d\mu = \int_A p d\mu = P(A)$$

[25, Thm. 5, p. 189]. Similarly $\lim_{n \rightarrow \infty} \tilde{Q}_n(A) = Q(A)$, so we find that $(\tilde{P}_n, \tilde{Q}_n) \rightarrow (P, Q)$, which completes the proof. \blacksquare

Theorem 19 (Decreasing). *Let $\mathcal{F} \supseteq \mathcal{F}_1 \supseteq \mathcal{F}_2 \supseteq \dots$ be a nonincreasing family of σ -algebras, and let $\mathcal{F}_\infty =$*

$\bigcap_{n=1}^\infty \mathcal{F}_n$ be the largest σ -algebra contained in all of them. Let $\alpha \in [0, \infty)$. If $\alpha \in [0, 1)$ or there exists an m such that $D_\alpha(P_{|\mathcal{F}_m} \| Q_{|\mathcal{F}_m}) < \infty$, then

$$\lim_{n \rightarrow \infty} D_\alpha(P_{|\mathcal{F}_n} \| Q_{|\mathcal{F}_n}) = D_\alpha(P_{|\mathcal{F}_\infty} \| Q_{|\mathcal{F}_\infty}).$$

The theorem cannot be extended to the case $\alpha = \infty$.

Lemma 5. *Let $\mathcal{F} \supseteq \mathcal{F}_1 \supseteq \mathcal{F}_2 \supseteq \dots$ be a nonincreasing family of σ -algebras. Let $\alpha \in (0, \infty)$, $p_n = \frac{dP_{|\mathcal{F}_n}}{d\mu_{|\mathcal{F}_n}}$, $q_n = \frac{dQ_{|\mathcal{F}_n}}{d\mu_{|\mathcal{F}_n}}$ and $X_n = f\left(\frac{p_n}{q_n}\right)$, where $f(x) = x^\alpha$ if $\alpha \neq 1$ and $f(x) = x \ln x + e^{-1}$ if $\alpha = 1$. If $\alpha \in (0, 1)$, or $\mathbf{E}_Q[X_1] < \infty$ and $P \ll Q$, then the family $\{X_n\}_{n \geq 1}$ is uniformly integrable (with respect to Q).*

Proof: Suppose first that $\alpha \in (0, 1)$. Then for any $b > 0$

$$\begin{aligned} \int_{X_n > b} X_n dQ &\leq \int_{X_n > b} X_n \left(\frac{X_n}{b}\right)^{(1-\alpha)/\alpha} dQ \\ &\leq b^{-(1-\alpha)/\alpha} \int X_n^{1/\alpha} dQ \leq b^{-(1-\alpha)/\alpha}, \end{aligned}$$

and, as $X_n \geq 0$, $\lim_{b \rightarrow \infty} \sup_n \int_{X_n > b} X_n dQ = 0$, which was to be shown. Alternatively, suppose that $\alpha \in [1, \infty)$ and assume without loss of generality that $\mathcal{F} = \mathcal{F}_1$. Then $\frac{p_n}{q_n} = \frac{dP_{|\mathcal{F}_n}}{dQ_{|\mathcal{F}_n}}$ (Q -a.s.) and hence by Proposition 1 and Jensen's inequality for conditional expectations

$$X_n = f\left(\mathbf{E}\left[\frac{dP}{dQ} \middle| \mathcal{F}_n\right]\right) \leq \mathbf{E}\left[f\left(\frac{dP}{dQ}\right) \middle| \mathcal{F}_n\right] = \mathbf{E}[X_1 | \mathcal{F}_n]$$

(Q -a.s.) As $\min_x x \ln x = -e^{-1}$, it follows that $X_n \geq 0$ and for any $b, c > 0$

$$\begin{aligned} \int_{|X_n| > b} |X_n| dQ &= \int_{X_n > b} X_n dQ \\ &\leq \int_{X_n > b} \mathbf{E}[X_1 | \mathcal{F}_n] dQ = \int_{X_n > b} X_1 dQ \\ &= \int_{X_n > b, X_1 \leq c} X_1 dQ + \int_{X_n > b, X_1 > c} X_1 dQ \\ &\leq c \cdot Q(X_n > b) + \int_{X_1 > c} X_1 dQ \\ &\leq \frac{c}{b} \mathbf{E}_Q[X_n] + \int_{X_1 > c} X_1 dQ \\ &\leq \frac{c}{b} \mathbf{E}_Q[X_1] + \int_{X_1 > c} X_1 dQ, \end{aligned}$$

where $\mathbf{E}_Q[X_n] \leq \mathbf{E}_Q[X_1]$ in the last inequality follows from the data processing inequality. Consequently,

$$\begin{aligned} \limsup_{b \rightarrow \infty} \int_{|X_n| > b} |X_n| dQ &= \lim_{c \rightarrow \infty} \limsup_{b \rightarrow \infty} \int_{|X_n| > b} |X_n| dQ \\ &\leq \lim_{c \rightarrow \infty} \lim_{b \rightarrow \infty} \frac{c}{b} \mathbf{E}_Q[X_1] + \lim_{c \rightarrow \infty} \int_{X_1 > c} X_1 dQ = 0, \end{aligned}$$

and the lemma follows. \blacksquare

Proof of Theorem 19: First suppose that $\alpha > 0$ and, for $n = 1, 2, \dots, \infty$, let $p_n = \frac{dP_{|\mathcal{F}_n}}{d\mu_{|\mathcal{F}_n}}$, $q_n = \frac{dQ_{|\mathcal{F}_n}}{d\mu_{|\mathcal{F}_n}}$ and $X_n = f\left(\frac{p_n}{q_n}\right)$ with $f(x) = x^\alpha$ if $\alpha \neq 1$ and $f(x) = x \ln x + e^{-1}$ if $\alpha = 1$, as in Lemma 5. If $\alpha \geq 1$, then assume without

loss of generality that $\mathcal{F} = \mathcal{F}_1$ and $m = 1$, such that $D_\alpha(P|_{\mathcal{F}_n} \| Q|_{\mathcal{F}_n}) < \infty$ implies $P \ll Q$. Now, for any $\alpha > 0$, it is sufficient to show that

$$\mathbf{E}_Q[X_n] \rightarrow \mathbf{E}_Q[X_\infty]. \quad (23)$$

By Proposition 1, $p_n = \mathbf{E}_\mu[p|\mathcal{F}_n]$ and $q_n = \mathbf{E}_\mu[q|\mathcal{F}_n]$. Therefore by a version of Lévy's theorem for decreasing sequences of σ -algebras [28, Theorem 6.23],

$$\begin{aligned} p_n &= \mathbf{E}_\mu[p|\mathcal{F}_n] \rightarrow \mathbf{E}_\mu[p|\mathcal{F}_\infty] = p_\infty, \\ q_n &= \mathbf{E}_\mu[q|\mathcal{F}_n] \rightarrow \mathbf{E}_\mu[q|\mathcal{F}_\infty] = q_\infty, \end{aligned} \quad (\mu\text{-a.s.})$$

and hence $X_n \rightarrow X_\infty$ (μ -a.s. and therefore Q -a.s.) If $0 < \alpha < 1$, then

$$\mathbf{E}_Q[X_n] = \mathbf{E}_\mu[p_n^\alpha q_n^{1-\alpha}] \leq \mathbf{E}_\mu[\alpha p_n + (1-\alpha)q_n] = 1 < \infty.$$

And if $\alpha \geq 1$, then by the data processing inequality $D_\alpha(P|_{\mathcal{F}_n} \| Q|_{\mathcal{F}_n}) < \infty$ for all n , which implies that also in this case $\mathbf{E}_Q[X_n] < \infty$. Hence uniform integrability (by Lemma 5) of the family of nonnegative random variables $\{X_n\}$ implies (23) [25, Thm. 5, p. 189], and the theorem follows for $\alpha > 0$. The remaining case, $\alpha = 0$, is proved by

$$\begin{aligned} \lim_{n \rightarrow \infty} D_0(P|_{\mathcal{F}_n} \| Q|_{\mathcal{F}_n}) &= \inf_n \inf_{\alpha > 0} D_\alpha(P|_{\mathcal{F}_n} \| Q|_{\mathcal{F}_n}) = \inf_{\alpha > 0} \inf_n D_\alpha(P|_{\mathcal{F}_n} \| Q|_{\mathcal{F}_n}) \\ &= \inf_{\alpha > 0} D_\alpha(P|_{\mathcal{F}_\infty} \| Q|_{\mathcal{F}_\infty}) = D_0(P|_{\mathcal{F}_\infty} \| Q|_{\mathcal{F}_\infty}). \end{aligned}$$

E. Distributions on Sequences

Suppose $(\mathcal{X}^\infty, \mathcal{F}^\infty)$ is the *direct product* of an infinite sequence of measurable spaces $(\mathcal{X}_1, \mathcal{F}_1), (\mathcal{X}_2, \mathcal{F}_2), \dots$. That is, $\mathcal{X}^\infty = \mathcal{X}_1 \times \mathcal{X}_2 \times \dots$ and \mathcal{F}^∞ is the smallest σ -algebra containing all the *cylinder sets*

$$S_n(A) = \{x^\infty \in \mathcal{X}^\infty \mid x_1, \dots, x_n \in A\}, \quad A \in \mathcal{F}^n,$$

for $n = 1, 2, \dots$, where $\mathcal{F}^n = \mathcal{F}_1 \otimes \dots \otimes \mathcal{F}_n$. Then a sequence of probability distributions P^1, P^2, \dots , where P^n is a distribution on $\mathcal{X}^n = \mathcal{X}_1 \times \dots \times \mathcal{X}_n$, is called *consistent* if

$$P^{n+1}(A \times \mathcal{X}_{n+1}) = P^n(A), \quad A \in \mathcal{F}^n.$$

For any such consistent sequence there exists a distribution P^∞ on $(\mathcal{X}^\infty, \mathcal{F}^\infty)$ such that its marginal distribution on \mathcal{X}^n is P^n , in the sense that

$$P^\infty(S_n(A)) = P^n(A), \quad A \in \mathcal{F}^n.$$

If P^1, P^2, \dots and Q^1, Q^2, \dots are two consistent sequences of probability distributions, then it is natural to ask whether the Rényi divergence $D_\alpha(P^n \| Q^n)$ converges to $D_\alpha(P^\infty \| Q^\infty)$. The following theorem shows that it does for $\alpha > 0$.

Theorem 20 (Consistent Distributions). *Let P^1, P^2, \dots and Q^1, Q^2, \dots be consistent sequences of probability distributions on $(\mathcal{X}^1, \mathcal{F}^1), (\mathcal{X}^2, \mathcal{F}^2), \dots$, where, for $n = 1, \dots, \infty$, $(\mathcal{X}^n, \mathcal{F}^n)$ is the direct product of the first n measurable spaces in the infinite sequence $(\mathcal{X}_1, \mathcal{F}_1), (\mathcal{X}_2, \mathcal{F}_2), \dots$. Then for any $\alpha \in (0, \infty]$*

$$D_\alpha(P^n \| Q^n) \rightarrow D_\alpha(P^\infty \| Q^\infty)$$

as $n \rightarrow \infty$.

Proof: Let $\mathcal{G}^n = \{S_n(A) \mid A \in \mathcal{F}^n\}$. Then

$$D_\alpha(P^n \| Q^n) = D_\alpha(P|_{\mathcal{G}^n} \| Q|_{\mathcal{G}^n}) \rightarrow D_\alpha(P^\infty \| Q^\infty)$$

by Theorem 18. ■

As a special case, we find that finite additivity of Rényi divergence, which is easy to verify, extends to countable additivity:

Theorem 21 (Additivity). *For $n = 1, 2, \dots$, let (P_n, Q_n) be pairs of probability distributions on measurable spaces $(\mathcal{X}_n, \mathcal{F}_n)$. Then for any $\alpha \in [0, \infty]$ and any $N \in \{1, 2, \dots\}$*

$$\sum_{n=1}^N D_\alpha(P_n \| Q_n) = D_\alpha(P_1 \times \dots \times P_N \| Q_1 \times \dots \times Q_N), \quad (24)$$

and, except for $\alpha = 0$, also

$$\sum_{n=1}^{\infty} D_\alpha(P_n \| Q_n) = D_\alpha(P_1 \times P_2 \times \dots \| Q_1 \times Q_2 \times \dots). \quad (25)$$

Countable additivity as in (25) does not hold for $\alpha = 0$. A counterexample is given following Example 2 below.

Proof: For simple orders α , (24) follows from independence of P_n and Q_n between different n , which implies that

$$\prod_{n=1}^N \int \left(\frac{dQ_n}{dP_n} \right)^{1-\alpha} dP_n = \int \left(\frac{d \prod_{n=1}^N Q_n}{d \prod_{n=1}^N P_n} \right)^{1-\alpha} d \prod_{n=1}^N P_n.$$

As N is finite, this extends to the extended orders by continuity in α . Finally, (25) follows from Theorem 20 by observing that the sequences $P^N = P_1 \times \dots \times P_N$ and $Q^N = Q_1 \times \dots \times Q_N$, for $N = 1, 2, \dots$, are consistent. ■

F. Absolute Continuity and Mutual Singularity

Shiryaev [25, pp. 366, 370] relates Hellinger integrals to absolute continuity and mutual singularity of probability distributions. His results may more elegantly be expressed in terms of Rényi divergence. They then follow from the observations that $D_0(P \| Q) = 0$ if and only if Q is absolutely continuous with respect to P and that $D_0(P \| Q) = \infty$ if and only if P and Q are mutually singular, together with right-continuity of $D_\alpha(P \| Q)$ in α at $\alpha = 0$.

Theorem 22 ([25, Theorem 2, p. 366]). *The following conditions are equivalent:*

- (i) $Q \ll P$,
- (ii) $Q(p > 0) = 1$,
- (iii) $D_0(P \| Q) = 0$,
- (iv) $\lim_{\alpha \downarrow 0} D_\alpha(P \| Q) = 0$.

Proof: Clearly (ii) is equivalent to $Q(p = 0) = 0$, which is equivalent to (i). The other cases follow by $\lim_{\alpha \downarrow 0} D_\alpha(P \| Q) = D_0(P \| Q) = -\ln Q(p > 0)$. ■

Theorem 23 ([25, Theorem 3, p. 366]). *The following conditions are equivalent:*

- (i) $P \perp Q$,
- (ii) $Q(p > 0) = 0$,
- (iii) $D_\alpha(P \| Q) = \infty$ for some $\alpha \in [0, 1)$,

(iv) $D_\alpha(P\|Q) = \infty$ for all $\alpha \in [0, \infty]$.

Proof: Equivalence of (i), (ii) and $D_0(P\|Q) = \infty$ follows from definitions. Equivalence of $D_0(P\|Q) = \infty$ and (iv) follows from the fact that Rényi divergence is continuous on $[0, 1]$ and nondecreasing in α . Finally, (iii) for some $\alpha \in (0, 1)$ is equivalent to

$$\int p^\alpha q^{1-\alpha} d\mu = 0,$$

which holds if and only if $pq = 0$ (μ -a.s.). It follows that in this case (iii) is equivalent to (i). ■

These properties give a convenient mathematical tool to prove absolute continuity or mutual singularity of infinite product distributions, as illustrated by the following proof by Shiryaev [25] of the *Gaussian dichotomy* [29]–[31].

Example 2 (Gaussian Dichotomy). Let $P = P_1 \times P_2 \times \dots$ and $Q = Q_1 \times Q_2 \times \dots$, where P_n and Q_n are Gaussian distributions with densities

$$p_n(x) = \frac{1}{\sqrt{\tau}} e^{-\frac{1}{2}(x-\mu_n)^2}, \quad q_n(x) = \frac{1}{\sqrt{\tau}} e^{-\frac{1}{2}(x-\nu_n)^2},$$

where $\tau = 2\pi$. Then

$$D_\alpha(P_n\|Q_n) = \frac{1}{2}\alpha(\mu_n - \nu_n)^2,$$

and by additivity for $\alpha > 0$

$$D_\alpha(P\|Q) = \frac{1}{2}\alpha \sum_{n=1}^{\infty} (\mu_n - \nu_n)^2.$$

Consequently, by Theorems 22 and 23 and symmetry in P and Q :

$$\begin{aligned} Q \ll P &\Leftrightarrow P \ll Q \Leftrightarrow \sum_{n=1}^{\infty} (\mu_n - \nu_n)^2 < \infty, \\ Q \perp P &\Leftrightarrow \sum_{n=1}^{\infty} (\mu_n - \nu_n)^2 = \infty. \end{aligned}$$

The observation that P and Q are either equivalent (both $P \ll Q$ and $Q \ll P$) or mutually singular is called the *Gaussian dichotomy*.

Example 2 shows that countable additivity does not hold for $\alpha = 0$: if $\sum_{n=1}^{\infty} (\mu_n - \nu_n)^2 = \infty$, then $\sum_{n=1}^N D_0(P_n\|Q_n) = 0$ for all N , while $D_0(P\|Q) = \infty$. In light of the proof of Theorem 21 this also provides a counterexample to (22) for $\alpha = 0$.

The Gaussian dichotomy raises the question of whether the same dichotomy holds for other product distributions. Let $P \sim Q$ denote that P and Q are *equivalent* (both $P \ll Q$ and $Q \ll P$). Suppose that $P = P_1 \times P_2 \times \dots$ and $Q = Q_1 \times Q_2 \times \dots$, where P_n and Q_n are arbitrary distributions on arbitrary measurable spaces. Then if $P_n \not\sim Q_n$ for some n , P and Q are not equivalent either. The question is therefore answered by the following theorem:

Theorem 24 (Kakutani's Dichotomy). *Let $\alpha \in (0, 1)$ and let $P = P_1 \times P_2 \times \dots$ and $Q = Q_1 \times Q_2 \times \dots$, where P_n and*

Q_n are distributions on arbitrary measurable spaces such that $P_n \sim Q_n$. Then

$$\begin{aligned} Q \sim P &\Leftrightarrow \sum_{n=1}^{\infty} D_\alpha(P_n\|Q_n) < \infty, \\ Q \perp P &\Leftrightarrow \sum_{n=1}^{\infty} D_\alpha(P_n\|Q_n) = \infty. \end{aligned}$$

Proof: If $\sum_{n=1}^{\infty} D_\alpha(P_n\|Q_n) = \infty$, then $D_\alpha(P\|Q) = \infty$ and $Q \perp P$ follows by Theorem 23. On the other hand, if $\sum_{n=1}^{\infty} D_\alpha(P_n\|Q_n) < \infty$, then for every $\varepsilon > 0$ there exists an N such that

$$\sum_{n=N+1}^{\infty} D_\alpha(P_n\|Q_n) \leq \varepsilon,$$

and consequently by additivity and monotonicity in α :

$$\begin{aligned} D_0(P\|Q) &= \lim_{\alpha \downarrow 0} D_\alpha(P\|Q) \\ &\leq \lim_{\alpha \downarrow 0} D_\alpha(P_1 \times \dots \times P_N\|Q_1 \times \dots \times Q_N) + \varepsilon = \varepsilon. \end{aligned}$$

As this holds for any $\varepsilon > 0$, $D_0(P\|Q)$ must equal 0, and, by Theorem 22, $Q \ll P$. As $Q \ll P$ implies $Q \not\perp P$, Theorem 23 implies that $D_\alpha(Q\|P) < \infty$, and by repeating the argument with the roles of P and Q reversed we find that also $P \ll Q$, which completes the proof. ■

Theorem 24 (with $\alpha = \frac{1}{2}$) is equivalent to a classical result by Kakutani [32], which was stated in terms of Hellinger integrals rather than Rényi divergence, and according to Gibbs and Su [22] might be responsible for popularising Hellinger integrals. As shown by Rényi [33], Kakutani's result is related to the amount of information that a sequence of observations contains about the parameter of a statistical model.

Contiguity and *entire separation* are asymptotic versions of absolute continuity and mutual singularity [34]. As might be expected, analogues of Theorems 22 and 23 also hold for these asymptotic concepts.

Let $(\mathcal{X}_n, \mathcal{F}_n)_{n=1,2,\dots}$ be a sequence of measurable spaces, and let $(P_n)_{n=1,2,\dots}$ and $(Q_n)_{n=1,2,\dots}$ be sequences of distributions on these spaces. Then the sequence (P_n) is *contiguous* with respect to the sequence (Q_n) , denoted $(P_n) \triangleleft (Q_n)$, if for all sequences of events $(A_n \in \mathcal{F}_n)_{n=1,2,\dots}$ such that $Q_n(A_n) \rightarrow 0$ as $n \rightarrow \infty$, we also have $P_n(A_n) \rightarrow 0$. If both $(P_n) \triangleleft (Q_n)$ and $(Q_n) \triangleleft (P_n)$, then the sequences are called *mutually contiguous* and we write $(P_n) \triangleleft\triangleright (Q_n)$. The sequences (P_n) and (Q_n) are *entirely separated*, denoted $(P_n) \triangle (Q_n)$, if there exist a sequence of events $(A_n \in \mathcal{F}_n)_{n=1,2,\dots}$ and a subsequence $(n_k)_{k=1,2,\dots}$ such that $P_{n_k}(A_{n_k}) \rightarrow 0$ and $Q_{n_k}(\mathcal{X}_{n_k} \setminus A_{n_k}) \rightarrow 0$ as $k \rightarrow \infty$.

Contiguity and entire separation are related to absolute continuity and mutual singularity in the following way [25, p. 369]: if $\mathcal{X}_n = \mathcal{X}$, $P_n = P$ and $Q_n = Q$ for all n , then

$$\begin{aligned} (P_n) \triangleleft (Q_n) &\Leftrightarrow P \ll Q, \\ (P_n) \triangleleft\triangleright (Q_n) &\Leftrightarrow P \sim Q, \\ (P_n) \triangle (Q_n) &\Leftrightarrow P \perp Q. \end{aligned}$$

Theorems 1 and 2 by Shiryaev [25, p. 370] imply the following two asymptotic analogues of Theorems 22 and 23:

Theorem 25. *The following conditions are equivalent:*

- (i) $(Q_n) \triangleleft (P_n)$,
- (ii) $\lim_{\alpha \downarrow 0} \limsup_{n \rightarrow \infty} D_\alpha(P_n \| Q_n) = 0$.

Theorem 26. *The following conditions are equivalent:*

- (i) $(P_n) \triangle (Q_n)$,
- (ii) $\lim_{\alpha \downarrow 0} \limsup_{n \rightarrow \infty} D_\alpha(P_n \| Q_n) = \infty$,
- (iii) $\limsup_{n \rightarrow \infty} D_\alpha(P_n \| Q_n) = \infty$ for some $\alpha \in (0, 1)$.
- (iv) $\limsup_{n \rightarrow \infty} D_\alpha(P_n \| Q_n) = \infty$ for all $\alpha \in (0, \infty]$.

G. Taylor Approximation for Parametric Models

Suppose $\{P_\theta \mid \theta \in \Theta \subseteq \mathbb{R}\}$ is a parametric statistical model. Then it is well known that, for sufficiently regular parametrisations, a second order Taylor approximation of $D(P_\theta \| P_{\theta'})$ in θ' at θ in the interior of Θ yields

$$\lim_{\theta' \rightarrow \theta} \frac{1}{(\theta - \theta')^2} D(P_\theta \| P_{\theta'}) = \frac{1}{2} J(\theta),$$

where $J(\theta) = \mathbf{E} \left[\left(\frac{d}{d\theta} \ln p_\theta \right)^2 \right]$ denotes the *Fisher information* at θ (see e.g. [27, Problem 12.7]). Haussler and Oppor [6] argue that this property generalises to

$$\lim_{\theta' \rightarrow \theta} \frac{1}{(\theta - \theta')^2} D_\alpha(P_\theta \| P_{\theta'}) = \frac{\alpha}{2} J(\theta)$$

for any $\alpha \in (0, \infty)$.

IV. MINIMAX RESULTS

A. Hypothesis Testing and Chernoff Information

Rényi divergence appears in bounds on the error probabilities when testing a probabilistic hypothesis Q against an alternative P [4], [35], [36]. This can be explained by the fact that $(1 - \alpha)D_\alpha(P \| Q)$ equals the *cumulant generating function* for the random variable $\ln(p/q)$ under the distribution Q (provided $\alpha \in (0, 1)$ or $P \ll Q$) [4]. The following theorem relates this cumulant generating function to two Kullback-Leibler divergences that involve the distribution P_α with density

$$p_\alpha = \frac{q^{1-\alpha} p^\alpha}{\int q^{1-\alpha} p^\alpha d\mu}, \quad (26)$$

which is well defined if and only if $0 < \int p^\alpha q^{1-\alpha} d\mu < \infty$.

Theorem 27. *For any simple order α*

$$(1 - \alpha)D_\alpha(P \| Q) = \inf_R \{ \alpha D(R \| P) + (1 - \alpha)D(R \| Q) \}, \quad (27)$$

with the convention that $\alpha D(R \| P) + (1 - \alpha)D(R \| Q) = \infty$ if it would otherwise be undefined. Moreover, if the distribution P_α with density (26) is well defined and $\alpha \in (0, 1)$ or $D(P_\alpha \| P) < \infty$, then the infimum is uniquely achieved by $R = P_\alpha$.

This result gives an interpretation of Rényi divergence as a trade-off between two Kullback-Leibler divergences.

Remark 1. Theorem 27 was formulated and proved for distributions on finite sets by Shayevitz [17], but appeared in the above formulation already in [16].

Proof of Theorem 27: First suppose that P_α is well defined or, equivalently, that $D_\alpha(P \| Q) < \infty$. Then for $\alpha \in (0, 1)$ or $D(R \| P) < \infty$, we have

$$\alpha D(R \| P) + (1 - \alpha)D(R \| Q) = D(R \| P_\alpha) - \ln \int p^\alpha q^{1-\alpha} d\mu.$$

Hence, if $0 < \alpha < 1$ or $D(P_\alpha \| P) < \infty$, the infimum over R is uniquely achieved by $R = P_\alpha$, for which it equals $(1 - \alpha)D_\alpha(P \| Q)$ as required. If, on the other hand, $\alpha > 1$ and $D(P_\alpha \| P) = \infty$, then we still have

$$\inf_R \{ \alpha D(R \| P) + (1 - \alpha)D(R \| Q) \} \geq (1 - \alpha)D_\alpha(P \| Q). \quad (28)$$

Secondly, suppose $\alpha \in (0, 1)$ and $D_\alpha(P \| Q) = \infty$. Then $P \perp Q$, and consequently either $D(R \| P) = \infty$ or $D(R \| Q) = \infty$ for all R , which means that (27) holds.

Next, consider the case that $\alpha > 1$ and $P \not\ll Q$. Then $D_\alpha(P \| Q) = \infty$ and the infimum over R is achieved by $R = P$, for which it equals $-\infty$, and again (27) holds.

Finally, we prove (27) for the remaining cases: $\alpha > 1$, $P \ll Q$ and either: (1) $D_\alpha(P \| Q) < \infty$, but $D(P_\alpha \| P) = \infty$; or (2) $D_\alpha(P \| Q) = \infty$. To this end, let $P_c = P(\cdot \mid p \leq cq)$ for all c that are sufficiently large that $P(p \leq cq) > 0$. The reader may verify that $D_\alpha(P_c \| Q) < \infty$ and $D(S \| P_c) < \infty$ for $s = p_c^\alpha q^{1-\alpha} / \int p_c^\alpha q^{1-\alpha} d\mu$, so that we have already proved that (27) holds if P is replaced by P_c . Hence, observing that for all R

$$D(R \| P_c) = \begin{cases} \infty & \text{if } R \not\ll P_c, \\ D(R \| P) + \ln P(p \leq cq) & \text{otherwise,} \end{cases}$$

we find that

$$\begin{aligned} \inf_R \{ \alpha D(R \| P) + (1 - \alpha)D(R \| Q) \} \\ \leq \limsup_{c \rightarrow \infty} \left(-\alpha \ln P(p \leq cq) \right. \\ \left. + \inf_R \{ \alpha D(R \| P_c) + (1 - \alpha)D(R \| Q) \} \right) \\ \leq \limsup_{c \rightarrow \infty} (1 - \alpha)D_\alpha(P_c \| Q) \leq (1 - \alpha)D_\alpha(P \| Q), \end{aligned}$$

where the last inequality follows by lower semi-continuity of D_α (Theorem 14). In case 2, (27) follows immediately. In case 1, (27) follows by combining this inequality with its converse (28). ■

Theorem 27 shows that $(1 - \alpha)D_\alpha(P \| Q)$ is the infimum over a set of functions that are linear in α , which implies the following corollary:

Corollary 2. *The function $(1 - \alpha)D_\alpha(P \| Q)$ is concave in α on $[0, \infty]$, with the conventions that it is 0 at $\alpha = 1$ even if $D(P \| Q) = \infty$ and that it is 0 at $\alpha = \infty$ if $P = Q$.*

Proof: Suppose first that $D(P \| Q) < \infty$. Then (27) also holds at $\alpha = 1$. Hence $(1 - \alpha)D_\alpha(P \| Q)$ is a point-wise infimum over linear functions on $(0, \infty)$, and thus concave. This extends to $\alpha \in \{0, \infty\}$ by continuity.

Alternatively, suppose that $D(P \| Q) = \infty$. Then $(1 - \alpha)D_\alpha(P \| Q)$ is still concave on $[0, 1]$, where it is also nonnegative. And by monotonicity of Rényi divergence, we have that $D_\alpha(P \| Q) = \infty$ for all $\alpha \geq 1$. Consequently,

$(1 - \alpha)D_\alpha(P\|Q)$ is nonnegative and concave for $\alpha \in [0, 1)$, at $\alpha = 1$ it is 0 (by convention) and for $\alpha \in (1, \infty]$ it is $-\infty$. It then follows that $(1 - \alpha)D_\alpha(P\|Q)$ is concave on all of $[0, \infty]$, as required. ■

As one might expect from continuity of $D_\alpha(P\|Q)$, the terms on the right-hand side of (27) are continuous in α , at least on $(0, 1)$:

Lemma 6. *If $D(P\|Q) < \infty$ or $D(Q\|P) < \infty$, then both $D(P_\alpha\|Q)$ and $D(P_\alpha\|P)$ are finite and continuous in α on $(0, 1)$.*

Proof: The lemma is symmetric in P and Q , so suppose without loss of generality that $D(P\|Q) < \infty$. Then $D_\alpha(P\|Q) \leq D(P\|Q) < \infty$ implies that P_α is well defined and finiteness of both $D(P_\alpha\|Q)$ and $D(P_\alpha\|P)$ follows from Theorem 27. Now observe that

$$D(P_\alpha\|Q) = \frac{1}{\int p^\alpha q^{1-\alpha} d\mu} \mathbf{E}_Q \left[\left(\frac{p}{q} \right)^\alpha \ln \left(\frac{p}{q} \right)^\alpha \right] + (1 - \alpha)D_\alpha(P\|Q).$$

Then by continuity of $D_\alpha(P\|Q)$ and hence of $\int p^\alpha q^{1-\alpha} d\mu$ in α , it is sufficient to verify continuity of $\mathbf{E}_Q[(p/q)^\alpha \ln(p/q)^\alpha]$. To this end, observe that

$$|(p/q)^\alpha \ln(p/q)^\alpha| \leq \begin{cases} 1/e & \text{if } p < q, \\ (p/q) \ln(p/q) & \text{if } p \geq q. \end{cases}$$

As $D(P\|Q) < \infty$ implies $\mathbf{E}_Q[\mathbf{1}_{\{p \geq q\}}(p/q) \ln(p/q)] < \infty$, we may apply the dominated convergence theorem to obtain

$$\lim_{\alpha \rightarrow \alpha^*} \mathbf{E}_Q \left[\left(\frac{p}{q} \right)^\alpha \ln \left(\frac{p}{q} \right)^\alpha \right] = \mathbf{E}_Q \left[\left(\frac{p}{q} \right)^{\alpha^*} \ln \left(\frac{p}{q} \right)^{\alpha^*} \right]$$

for any $\alpha^* \in (0, 1)$, which proves continuity of $D(P_\alpha\|Q)$. Continuity of $D(P_\alpha\|P)$ now follows from Theorem 27 and continuity of $(1 - \alpha)D_\alpha(P\|Q)$. ■

Theorem 28. *Suppose that $D(P\|Q) < \infty$. Then the following minimax identity holds:*

$$\sup_{\alpha \in (0, \infty)} \inf_R \{ \alpha D(R\|P) + (1 - \alpha)D(R\|Q) \} = \inf_R \sup_{\alpha \in (0, \infty)} \{ \alpha D(R\|P) + (1 - \alpha)D(R\|Q) \}, \quad (29)$$

with the convention that $\alpha D(R\|P) + (1 - \alpha)D(R\|Q) = \infty$ if it would otherwise be undefined. Moreover, (29) still holds if α is restricted to $(0, 1)$ on its left-hand side, and if there exists an $\alpha^* \in (0, 1)$ such that $D(P_{\alpha^*}\|P) = D(P_{\alpha^*}\|Q)$, then (α^*, P_{α^*}) is a saddle-point for (29) and both sides of (29) are equal to

$$(1 - \alpha^*)D_{\alpha^*}(P\|Q) = \sup_{\alpha \in (0, 1)} (1 - \alpha)D_\alpha(P\|Q) = D(P_{\alpha^*}\|P) = D(P_{\alpha^*}\|Q). \quad (30)$$

The minimax value defined in (29) is the *Chernoff information*, which gives an asymptotically tight bound on both the type 1 and the type 2 errors in tests of P vs. Q . The same connection between Chernoff information and $D(P_{\alpha^*}\|P)$ is discussed by Cover and Thomas [27, Section 12.9], with a different proof.

Proof of Theorem 28: Let $f(\alpha, R) = \alpha D(R\|P) + (1 - \alpha)D(R\|Q)$. For $\alpha \in (0, 1)$, $D_\alpha(P\|Q) \leq D(P\|Q) < \infty$ implies that P_α is well defined. Suppose there exists $\alpha^* \in (0, 1)$ such that $D(P_{\alpha^*}\|P) = D(P_{\alpha^*}\|Q)$. Then Theorem 27 implies that (α^*, P_{α^*}) is a saddle-point for $f(\alpha, R)$, so that (29) holds [37, Lemma 36.2], and Theorem 27 also implies that all quantities in (30) are equal to $f(\alpha^*, P_{\alpha^*})$.

Let \mathcal{A} be either $(0, 1)$ or $(0, \infty)$. As the sup inf is never bigger than the inf sup [37, Lemma 36.1], we have that

$$\sup_{\alpha \in \mathcal{A}} \inf_R f(\alpha, R) \leq \sup_{\alpha \in (0, \infty)} \inf_R f(\alpha, R) \leq \inf_R \sup_{\alpha \in (0, \infty)} f(\alpha, R),$$

so it remains to prove the converse inequality.

By Lemma 6 we know that both $D(P_\alpha\|P)$ and $D(P_\alpha\|Q)$ are finite and continuous in α on $(0, 1)$. By the intermediate value theorem, there are therefore three possibilities: (1) there exists $\alpha^* \in (0, 1)$ such that $D(P_{\alpha^*}\|P) = D(P_{\alpha^*}\|Q)$, for which we have already proved (29); (2) $D(P_\alpha\|P) < D(P_\alpha\|Q)$ for all $\alpha \in (0, 1)$; and (3) $D(P_\alpha\|P) > D(P_\alpha\|Q)$ for all $\alpha \in (0, 1)$.

We proceed with case (2), observing that

$$\begin{aligned} \inf_R \sup_{\alpha \in (0, \infty)} f(\alpha, R) &= \inf_{R: D(R\|Q) < \infty} \sup_{\alpha \in (0, \infty)} f(\alpha, R) \\ &= \inf_{R: D(R\|Q) < \infty} \left\{ D(R\|Q) \right. \\ &\quad \left. + \sup_{\alpha \in (0, \infty)} \alpha (D(R\|P) - D(R\|Q)) \right\} \\ &= \inf_{R: D(R\|P) \leq D(R\|Q) < \infty} D(R\|Q) \\ &\leq \inf_{0 < \alpha < 1} D(P_\alpha\|Q). \end{aligned}$$

Now by Theorem 27

$$\begin{aligned} \inf_{0 < \alpha < 1} D(P_\alpha\|Q) &\leq \liminf_{\alpha \downarrow 0} D(P_\alpha\|Q) \\ &= \liminf_{\alpha \downarrow 0} \left\{ D_\alpha(P\|Q) - \frac{\alpha}{1 - \alpha} D(P_\alpha\|P) \right\} \\ &\leq \liminf_{\alpha \downarrow 0} D_\alpha(P\|Q) = \liminf_{\alpha \downarrow 0} (1 - \alpha)D_\alpha(P\|Q) \\ &= \liminf_{\alpha \downarrow 0} f(\alpha, R) \leq \sup_{\alpha \in \mathcal{A}} \inf_R f(\alpha, R), \end{aligned}$$

as required. It remains to consider case (3), which turns out to be impossible by the following argument: two applications of Theorem 27 give

$$\begin{aligned} D_{1/2}(P\|Q) &= \inf_{0 < \alpha < 1} \left\{ D(P_\alpha\|P) + D(P_\alpha\|Q) \right\} \\ &\leq 2 \inf_{0 < \alpha < 1} D(P_\alpha\|P) \leq 2 \limsup_{\alpha \uparrow 1} D(P_\alpha\|P) \\ &= 2 \limsup_{\alpha \uparrow 1} \left\{ \frac{1 - \alpha}{\alpha} D_\alpha(P\|Q) - \frac{1 - \alpha}{\alpha} D(P_\alpha\|P) \right\} \\ &\leq 2 \limsup_{\alpha \uparrow 1} \frac{1 - \alpha}{\alpha} D_\alpha(P\|Q) = 0. \end{aligned}$$

It follows that $P = Q$, which contradicts the assumption that $D(P_\alpha\|P) > D(P_\alpha\|Q)$ for any $\alpha \in (0, 1)$. ■

B. Channel Capacity and Minimax Redundancy

Consider a non-empty family $\{P_\theta \mid \theta \in \Theta\}$ of probability distributions on a sample space \mathcal{X} . We may think of θ as

a parameter in a statistical model or as an input letter of an information channel. In the main results of this section we will only consider finite \mathcal{X} , with n elements. Whenever distributions on Θ are involved, we also implicitly assume that Θ is a topological space that is equipped with the Borel σ -algebra, and that the map $\theta \mapsto P_\theta$ is measurable.

We will study

$$C_\alpha = \sup_{\pi} \inf_Q \int D_\alpha(P_\theta \| Q) d\pi(\theta),$$

which has been proposed as the appropriate generalisation of the *channel capacity* from $\alpha = 1$ to general α [4], [18].

If \mathcal{X} is finite, then the channel capacity is also finite:

Theorem 29. *If \mathcal{X} has n elements, then $C_\alpha \leq \ln n$ for any $\alpha \in [0, \infty]$.*

Proof: Let U denote the uniform distribution on \mathcal{X} . Then

$$\begin{aligned} \sup_{\pi} \inf_Q \int D_\alpha(P_\theta \| Q) d\pi(\theta) &\leq \sup_{\pi} \int D_\alpha(P_\theta \| U) d\pi(\theta) \\ &= \sup_{\theta} D_\alpha(P_\theta \| U) \leq \sup_{\theta} D_\infty(P_\theta \| U) \\ &= \sup_{\theta} \ln \max_x \frac{P_\theta(x)}{1/n} \leq \ln n. \end{aligned}$$

For $\alpha = 1$, it is a classical result by Gallager and Ryabko [38] that the channel capacity equals the *minimax redundancy*:

$$R_\alpha = \inf_Q \sup_{\theta \in \Theta} D_\alpha(P_\theta \| Q).$$

For finite Θ , Csiszár [4] has shown that this result in fact extends to any $\alpha \in (0, \infty)$, noting that minimax redundancy R_α (and therefore the channel capacity C_α) may be geometrically interpreted as the “radius” of the family of distributions $\{P_\theta \mid \theta \in \Theta\}$ with respect to the Rényi divergence of order α . It turns out that Csiszár’s result extends to general Θ and all orders α :

Theorem 30. *Suppose \mathcal{X} is finite. Then for any $\alpha \in [0, \infty]$ the channel capacity equals the minimax redundancy:*

$$C_\alpha = R_\alpha. \quad (31)$$

For $\alpha = 1$, Haussler [39] has extended this result to infinite sample spaces \mathcal{X} . It seems plausible that his approach might extend to other orders α as well.

Equation 31 is equivalent to the minimax identity

$$\sup_{\pi} \inf_Q \psi_\alpha(\pi, Q) = \inf_Q \sup_{\pi} \psi_\alpha(\pi, Q), \quad (32)$$

where

$$\psi_\alpha(\pi, Q) = \int D_\alpha(P_\theta \| Q) d\pi(\theta).$$

We will prove this identity using Sion’s minimax theorem [40], [41], which we state with its arguments exchanged to make them line up with the arguments of ψ_α :

Theorem 31 (Sion’s Minimax Theorem). *Let A be a convex subset of a linear topological space and B a compact convex subset of a linear topological space. Let $f: A \times B \rightarrow \mathbb{R}$ be such that*

- (i) *$f(\cdot, b)$ is upper semi-continuous and quasi-concave on A for each $b \in B$;*
- (ii) *$f(a, \cdot)$ is lower semi-continuous and quasi-convex on B for each $a \in A$.*

Then

$$\sup_{a \in A} \min_{b \in B} f(a, b) = \min_{b \in B} \sup_{a \in A} f(a, b).$$

Proof of Theorem 30: Sion’s minimax theorem cannot be applied directly, because ψ_α may be infinite. For $\lambda \in (0, 1)$, we therefore introduce the auxiliary function

$$\psi_\alpha^\lambda(\pi, Q) = \psi_\alpha(\pi, (1 - \lambda)U + \lambda Q),$$

where U is the uniform distribution on \mathcal{X} . Finiteness of ψ_α^λ follows from

$$\begin{aligned} D_\alpha(P_\theta \| (1 - \lambda)U + \lambda Q) &\leq D_\alpha(P_\theta \| U) - \ln(1 - \lambda) \\ &\leq D_\infty(P_\theta \| U) - \ln(1 - \lambda) \leq \ln n - \ln(1 - \lambda), \end{aligned} \quad (33)$$

where n denotes the number of elements in \mathcal{X} .

To verify the other conditions of Theorem 31, we observe that $\psi_\alpha^\lambda(\cdot, Q)$ is linear, and hence continuous and concave. Convexity of $\psi_\alpha^\lambda(\pi, \cdot)$ follows from convexity of $\psi_\alpha(\pi, \cdot)$, which holds because $\psi_\alpha(\pi, \cdot)$ is a linear combination of convex functions. Continuity of $\psi_\alpha^\lambda(\pi, \cdot)$ follows by the dominated convergence theorem (which applies by (33)) and continuity of $D_\alpha(P_\theta \| \cdot)$. Thus we may apply Sion’s minimax theorem.

By

$$D_\alpha(P_\theta \| (1 - \lambda)U + \lambda Q) \leq D_\alpha(P_\theta \| Q) - \ln \lambda,$$

we also have $\psi_\alpha^\lambda(\pi, Q) \leq \psi_\alpha(\pi, Q) - \ln \lambda$, and hence we may reason as follows:

$$\begin{aligned} \sup_{\pi} \inf_Q \psi_\alpha(\pi, Q) - \ln \lambda &\geq \sup_{\pi} \inf_Q \psi_\alpha^\lambda(\pi, Q) \\ &= \inf_Q \sup_{\pi} \psi_\alpha^\lambda(\pi, Q) \geq \inf_Q \sup_{\pi} \psi_\alpha(\pi, Q). \end{aligned}$$

By letting λ tend to 1 we find

$$\sup_{\pi} \inf_Q \psi_\alpha(\pi, Q) \geq \inf_Q \sup_{\pi} \psi_\alpha(\pi, Q).$$

As the $\sup \inf$ never exceeds the $\inf \sup$ [37, Lemma 36.1], the converse inequality also holds, and the proof is complete. ■

A distribution π_{opt} on the parameter space Θ is a *capacity achieving* input distribution if

$$\inf_Q \int D_\alpha(P_\theta \| Q) d\pi_{\text{opt}}(\theta) = C_\alpha.$$

A distribution Q_{opt} on \mathcal{X} may be called a *redundancy achieving* distribution if

$$\sup_{\theta} D_\alpha(P_\theta \| Q_{\text{opt}}) = R_\alpha.$$

If the sample space is finite, then a redundancy achieving distribution always exists:

Lemma 7. *Suppose \mathcal{X} is finite and let $\alpha \in [0, \infty]$. Then the function $Q \mapsto \sup_{\theta} D_\alpha(P_\theta \| Q)$ is continuous and convex, and has at least one minimum. Consequently, a redundancy achieving distribution Q_{opt} exists.*

For $\alpha > 0$, the redundancy achieving distribution is in fact unique, as will be shown by Theorem 34 below.

Proof: Denote the number of elements in \mathcal{X} by n , let $\Delta_n = \{(p_1, \dots, p_n) \mid \sum_{i=1}^n p_i = 1, p_i \geq 0\}$ denote the probability simplex on n outcomes, and let $f(Q) = \sup_{\theta} D_{\alpha}(P_{\theta} \| Q)$. Then the domain of f is Δ_n , and since f is the supremum over continuous, convex functions, it is lower semi-continuous and convex itself. As convexity on a simplex implies upper semi-continuity [37, Theorem 10.2], it follows that f is both lower and upper semi-continuous, and is therefore continuous. As the domain of f is compact, this implies that it also attains its minimum. ■

Theorem 32. *Suppose \mathcal{X} is finite and let $\alpha \in [0, \infty]$. If there exists a (possibly non-unique) capacity achieving input distribution π_{opt} , then $\int D_{\alpha}(P_{\theta} \| Q) d\pi_{\text{opt}}(\theta)$ is minimized by $Q = Q_{\text{opt}}$ and $D_{\alpha}(P_{\theta} \| Q_{\text{opt}}) = R_{\alpha}$ almost surely under π_{opt} .*

If R_{α} is regarded as the radius of $\{P_{\theta} \mid \theta \in \Theta\}$, then this theorem shows how Q_{opt} may be interpreted as its center.

Proof: Since π_{opt} is capacity achieving,

$$\begin{aligned} C_{\alpha} &= \inf_Q \int D_{\alpha}(P_{\theta} \| Q) d\pi_{\text{opt}}(\theta) \\ &\leq \int D_{\alpha}(P_{\theta} \| Q_{\text{opt}}) d\pi_{\text{opt}}(\theta) \\ &\leq \int R_{\alpha} d\pi_{\text{opt}}(\theta) = R_{\alpha} = C_{\alpha}. \end{aligned}$$

The result follows because both inequalities must be equalities. ■

Three orders α for the channel capacity C_{α} and minimax redundancy R_{α} are of particular interest. The classical ones are $\alpha = 1$, because it corresponds to the original definition of channel capacity by Shannon, and $\alpha = 0$ because C_0 gives an upper bound on the zero error capacity, which also dates back to Shannon.

Now let us look at the case $\alpha = \infty$, assuming for simplicity that \mathcal{X} is finite. We find that

$$\begin{aligned} \sup_{\theta} D_{\infty}(P_{\theta} \| Q) &= \sup_{\theta} \max_x \ln \frac{P_{\theta}(x)}{Q(x)} \\ &= \max_x \ln \frac{\sup_{\theta} P_{\theta}(x)}{Q(x)} \end{aligned}$$

is the *worst-case regret* of Q relative to $\{P_{\theta} \mid \theta \in \Theta\}$ [3]. It is well known [3], [42] that the distribution that minimizes the worst-case regret is uniquely given by the *normalized maximum likelihood* or *Shtarkov distribution*

$$S(x) = \frac{\sup_{\theta} P_{\theta}(x)}{\sum_x \sup_{\theta} P_{\theta}(x)},$$

which achieves worst-case regret

$$R_{\infty} = \sum_x \sup_{\theta} P_{\theta}(x).$$

Thus in this case $Q_{\text{opt}} = S$ is unique. Moreover, by some algebra we can quantify the amount by which any other distribution $Q \neq S$ exceeds the minimax redundancy, namely by $D_{\infty}(S \| Q)$:

Theorem 33. *Suppose \mathcal{X} is finite. Then the worst-case regret of any distribution Q satisfies*

$$\sup_{\theta} D_{\infty}(P_{\theta} \| Q) = \sup_{\theta} D_{\infty}(P_{\theta} \| S) + D_{\infty}(S \| Q). \quad (34)$$

Proof: We have

$$\begin{aligned} \max_x \ln \frac{\sup_{\theta} P_{\theta}(x)}{Q(x)} &= \max_x \left(\ln \frac{\sup_{\theta} P_{\theta}(x)}{S(x)} + \ln \frac{S(x)}{Q(x)} \right) \\ &= \ln \sum_x \sup_{\theta} P_{\theta}(x) + \max_x \ln \frac{S(x)}{Q(x)} \\ &= \max_x \ln \frac{\sup_{\theta} P_{\theta}(x)}{S(x)} + \max_x \ln \frac{S(x)}{Q(x)}. \end{aligned}$$

The previous result generalises to any positive order α as a one-sided inequality:

Theorem 34. *Suppose \mathcal{X} is finite. Then, for $\alpha \in (0, \infty]$,*

$$Q_{\text{opt}} = \arg \min_Q \sup_{\theta} D_{\alpha}(P_{\theta} \| Q)$$

uniquely exists and for all Q

$$\sup_{\theta} D_{\alpha}(P_{\theta} \| Q) \geq \sup_{\theta} D_{\alpha}(P_{\theta} \| Q_{\text{opt}}) + D_{\alpha}(Q_{\text{opt}} \| Q). \quad (35)$$

This result, which is new, is reminiscent of Sibson's identity [4], [43]. It shows that any distribution Q that is close to achieving the minimax redundancy in the sense that

$$\sup_{\theta} D_{\alpha}(P_{\theta} \| Q) \leq \sup_{\theta} D_{\alpha}(P_{\theta} \| Q_{\text{opt}}) + \delta,$$

must be close to Q_{opt} in the sense that

$$D_{\alpha}(Q_{\text{opt}} \| Q) \leq \delta.$$

As shown in Example 3 below, Theorem 34 cannot be extended to $\alpha = 0$. For $\alpha > 0$, we will prove it by expressing it as a minimax identity for the function

$$\phi_{\alpha}(R, Q) = \sup_{\theta \in \Theta} D_{\alpha}(P_{\theta} \| Q) - D_{\alpha}(R \| Q),$$

where we adopt the convention that $\phi_{\alpha}(R, Q) = \infty$ if both $\sup_{\theta \in \Theta} D_{\alpha}(P_{\theta} \| Q)$ and $D_{\alpha}(R \| Q)$ are infinite.

Lemma 8. *Suppose \mathcal{X} is finite. Then, for $\alpha \in (0, \infty]$,*

$$\max_R \min_Q \phi_{\alpha}(R, Q) = \min_Q \max_R \phi_{\alpha}(R, Q). \quad (36)$$

Moreover, $Q_{\text{opt}} = \arg \min_Q \sup_{\theta \in \Theta} D_{\alpha}(P_{\theta} \| Q)$ uniquely exists and $(R, Q) = (Q_{\text{opt}}, Q_{\text{opt}})$ is a saddle-point.

Theorem 34 then follows by the following argument.

Proof of Theorem 34: The definition of a saddle-point implies that $\phi_{\alpha}(Q_{\text{opt}}, Q) \geq \phi_{\alpha}(Q_{\text{opt}}, Q_{\text{opt}})$ for all Q , from which the theorem follows after plugging in the definition of ϕ_{α} . ■

To prove Lemma 8, we will use Sion's minimax theorem (Theorem 31) again. Verifying its conditions requires the following lemmas.

Lemma 9. *For any $\alpha \in [0, \infty]$, $\phi_{\alpha}(\cdot, Q)$ is quasi-concave for any Q and $\phi_{\alpha}(R, \cdot)$ is quasi-convex for any R .*

Proof: Quasi-concavity in the first argument follows because $D_\alpha(R\|Q)$ is quasi-convex in R . To show quasi-convexity in the second argument, let R, Q_0, Q_1 and $\lambda \in (0, 1)$ be arbitrary, define $Q_\lambda = (1 - \lambda)Q_0 + \lambda Q_1$, and observe that

$$\begin{aligned} \phi_\alpha(R, Q_\lambda) &= \sup_\theta D_\alpha(P_\theta\|Q_\lambda) - D_\alpha(R\|Q_\lambda) \\ &\leq \sup_\theta \left((1 - \lambda)D_\alpha(P_\theta\|Q_0) + \lambda D_\alpha(P_\theta\|Q_1) \right) - D_\alpha(R\|Q_\lambda) \\ &\leq (1 - \lambda) \sup_\theta D_\alpha(P_\theta\|Q_0) + \lambda \sup_\theta D_\alpha(P_\theta\|Q_1) \\ &\quad - D_\alpha(R\|Q_\lambda). \end{aligned}$$

As $D_\alpha(R\|Q_\lambda)$ is convex in λ and the other terms are linear, this upper bound is concave in λ . It follows that it is maximized at one of the endpoints of its domain, $\lambda = 0$ and $\lambda = 1$, where it equals $\phi_\alpha(R, Q_0)$ or $\phi_\alpha(R, Q_1)$, respectively. We therefore find that

$$\phi_\alpha(R, Q_\lambda) \leq \max\{\phi_\alpha(R, Q_0), \phi_\alpha(R, Q_1)\},$$

which was to be shown. \blacksquare

Lemma 10. *Let $\alpha \in (0, \infty]$. Then $\phi_\alpha(\cdot, Q)$ is upper semi-continuous for any Q (in the topology of setwise convergence).*

Proof: By lower semi-continuity of $D_\alpha(\cdot\|\cdot)$. \blacksquare

Let $\Delta_n = \{(p_1, \dots, p_n) \mid \sum_{i=1}^n p_i = 1, p_i \geq 0\}$ denote the probability simplex on n outcomes, and let $\text{ri}(\Delta_n) = \{(p_1, \dots, p_n) \mid \sum_{i=1}^n p_i = 1, p_i > 0\}$ denote its relative interior in \mathbb{R}^n .

Lemma 11. *Suppose \mathcal{X} has n elements. Then, for $\alpha \in [0, \infty]$ and any R , $\phi_\alpha(R, \cdot)$ is continuous on $\text{ri}(\Delta_n)$ and upper semi-continuous on Δ_n .*

Proof: The function $\sup_\theta D_\alpha(P_\theta\|\cdot)$ is continuous on Δ_n by Lemma 7, and $D_\alpha(R\|\cdot)$ is continuous by Theorem 17. It follows that their difference is continuous as long as at least one of the two is finite.

On $\text{ri}(\Delta_n)$ both are finite, and hence $\phi_\alpha(R, \cdot)$ is continuous. Only for a sequence Q_1, Q_2, \dots converging to a point Q^* on the boundary of Δ_n may continuity break down, if both $\sup_\theta D_\alpha(P_\theta\|Q^*)$ and $D_\alpha(R\|Q^*)$ are infinite. In this case $\phi_\alpha(R, Q^*) = \infty$ by definition, and hence we still have upper semi-continuity. \blacksquare

Proof of Lemma 8: Let n denote the number of elements in \mathcal{X} and, for $\epsilon \in (0, 1/n]$, define $\Delta_n^\epsilon = \{(p_1, \dots, p_n) \mid \sum_{i=1}^n p_i = 1, p_i \geq \epsilon\}$. Then, on $\Delta_n \times \Delta_n^\epsilon$, Lemmas 9 through 11 show that $\phi_\alpha(R, Q)$ is upper semi-continuous and quasi-concave in R , and continuous and quasi-convex in Q . Thus it satisfies the conditions of Theorem 31, so that

$$\max_R \min_{Q \in \Delta_n^\epsilon} \phi_\alpha(R, Q) = \min_{Q \in \Delta_n^\epsilon} \max_R \phi_\alpha(R, Q). \quad (37)$$

(The suprema over R are attained, because $\phi_\alpha(\cdot, Q)$ and hence also $\min_{Q \in \Delta_n^\epsilon} \phi_\alpha(\cdot, Q)$ are upper semi-continuous functions on a compact domain.)

Let $Q_{\text{opt}}^\epsilon \in \arg \min_{Q \in \Delta_n^\epsilon} \sup_\theta D_\alpha(P_\theta\|Q)$ be a distribution that achieves the minimum on the right-hand side of (37). Then

for any $R \neq Q_{\text{opt}}^\epsilon$ we have

$$\begin{aligned} \min_{Q \in \Delta_n^\epsilon} \phi_\alpha(R, Q) &\leq \phi_\alpha(R, Q_{\text{opt}}^\epsilon) \\ &< \sup_\theta D_\alpha(P_\theta\|Q_{\text{opt}}^\epsilon) = \min_{Q \in \Delta_n^\epsilon} \max_R \phi_\alpha(R, Q), \end{aligned}$$

so only $R = Q_{\text{opt}}^\epsilon$ can achieve the maximum on the left-hand side of (37). It follows that $(R, Q) = (Q_{\text{opt}}^\epsilon, Q_{\text{opt}}^\epsilon)$ is a saddle-point, and that Q_{opt}^ϵ is unique.

Let $\epsilon_1 > \epsilon_2 > \dots > 0$ be a decreasing sequence that converges to 0. Then $Q_{\text{opt}}^{\epsilon_1}, Q_{\text{opt}}^{\epsilon_2}, \dots$ is an infinite sequence in a compact domain, and hence (by the Bolzano-Weierstrass theorem) there is a subsequence $\epsilon'_1 > \epsilon'_2 > \dots > 0$ such that $Q_{\text{opt}}^{\epsilon'_1}, Q_{\text{opt}}^{\epsilon'_2}, \dots$ converges to some $Q^* \in \Delta_n$.

Now let $Q \in \text{ri}(\Delta_n)$ be arbitrary. Then upper semi-continuity of $\phi_\alpha(\cdot, Q)$ implies that

$$\begin{aligned} \phi_\alpha(Q^*, Q) &\geq \limsup_{m \rightarrow \infty} \phi_\alpha(Q_{\text{opt}}^{\epsilon'_m}, Q) \\ &\geq \limsup_{m \rightarrow \infty} \min_{Q \in \Delta_n^{\epsilon'_m}} \phi_\alpha(Q_{\text{opt}}^{\epsilon'_m}, Q) \\ &= \limsup_{m \rightarrow \infty} \min_{Q \in \Delta_n^{\epsilon'_m}} \max_R \phi_\alpha(R, Q) \\ &\geq \inf_{Q \in \Delta_n} \max_R \phi_\alpha(R, Q). \end{aligned}$$

Together with upper semi-continuity of $\phi_\alpha(Q^*, \cdot)$ on Δ_n (see Lemma 11) this implies that also for any Q on the boundary of Δ_n

$$\begin{aligned} \phi_\alpha(Q^*, Q) &\geq \limsup_{\lambda \uparrow 1} \phi_\alpha(Q^*, (1 - \lambda)U + \lambda Q) \\ &\geq \inf_{Q \in \Delta_n} \max_R \phi_\alpha(R, Q), \end{aligned}$$

where $U = (1/n, \dots, 1/n)$ is the uniform distribution. Thus

$$\begin{aligned} \max_R \inf_{Q \in \Delta_n} \phi_\alpha(R, Q) &\geq \inf_{Q \in \Delta_n} \phi_\alpha(Q^*, Q) \\ &\geq \inf_{Q \in \Delta_n} \max_R \phi_\alpha(R, Q). \end{aligned} \quad (38)$$

Since the $\max \inf$ never exceeds the $\inf \max$ [37, Lemma 36.1], these inequalities must in fact hold with equality.

It remains to establish that Q_{opt} uniquely exists and that $(R, Q) = (Q_{\text{opt}}, Q_{\text{opt}})$ is a saddle-point. By Lemma 7 we know that there exists a distribution Q' that minimizes $\sup_\theta D_\alpha(P_\theta\|\cdot)$. Suppose $Q' \neq Q^*$. Then $\phi_\alpha(Q^*, Q') < \min_Q \sup_\theta D_\alpha(P_\theta\|Q)$, which would contradict (38). Hence $Q_{\text{opt}} = Q' = Q^*$ is unique.

Moreover, (38) implies that

$$\phi_\alpha(Q_{\text{opt}}, Q) \geq \min_Q \max_R \phi_\alpha(R, Q) = \phi_\alpha(Q_{\text{opt}}, Q_{\text{opt}})$$

for all Q , and it may be directly verified that $\phi_\alpha(R, Q_{\text{opt}}) \leq \phi_\alpha(Q_{\text{opt}}, Q_{\text{opt}})$ for all R . Thus $(Q_{\text{opt}}, Q_{\text{opt}})$ is a saddle-point, which concludes the proof. \blacksquare

A distribution π on the parameter space Θ is called a *barycentric input distribution* if

$$Q_{\text{opt}} = \int P_\theta d\pi(\theta).$$

Example 3. Take $\alpha \in (0, \infty]$ and consider the distributions

$$P_1 = (\frac{1}{2}, 0, \frac{1}{2}), \quad P_2 = (0, \frac{1}{2}, \frac{1}{2})$$

on a three-element set. Then by symmetry the unique redundancy achieving distribution has the form

$$Q_{\text{opt}} = (q, q, 1 - 2q).$$

If α is a simple order, then for $\theta \in \{1, 2\}$ the divergence is

$$\begin{aligned} D_\alpha(P_\theta \| Q_{\text{opt}}) &= \frac{1}{\alpha - 1} \ln \left(\left(\frac{1}{2}\right)^\alpha q^{1-\alpha} + \left(\frac{1}{2}\right)^\alpha (1 - 2q)^{1-\alpha} \right) \\ &= \frac{\alpha \ln 2}{1 - \alpha} + \frac{1}{\alpha - 1} \ln (q^{1-\alpha} + (1 - 2q)^{1-\alpha}). \end{aligned}$$

To find q , we therefore we have to extremize

$$f(q) = q^{1-\alpha} + (1 - 2q)^{1-\alpha},$$

which leads to

$$q = \frac{1}{2 + 2^{\frac{1}{\alpha}}}. \quad (39)$$

The reader may verify that (39) also holds for $\alpha = 1$, giving $Q_{\text{opt}} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{2})$, and for $\alpha = \infty$, giving $Q_{\text{opt}} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Note that only for $\alpha = 1$ is Q_{opt} a convex combination of P_1 and P_2 , with unique barycentric input distribution $\pi = (\frac{1}{2}, \frac{1}{2})$.

Finally, consider $\alpha = 0$, for which Theorem 34 does not apply. In this case (39) still holds, giving $Q_{\text{opt}} = (0, 0, 1)$. Now let $Q = (\frac{1}{2}, \frac{1}{2}, 0)$. Then, for $\theta \in \{1, 2\}$, we see that the first two terms in (35) are well-behaved:

$$\begin{aligned} \limsup_{\alpha \downarrow 0} D_\alpha(P_\theta \| Q) &= \sup_{\theta} D_0(P_\theta \| Q) = \ln 2, \\ \limsup_{\alpha \downarrow 0} D_\alpha(P_\theta \| Q_{\text{opt}}) &= 0 = \sup_{\theta} D_0(P_\theta \| Q_{\text{opt}}). \end{aligned}$$

For the last term, however, $\lim_{\alpha \downarrow 0} D_\alpha(Q_{\text{opt}} \| Q) = \ln 2$, whereas $D_0(Q_{\text{opt}} \| Q) = \infty$, and so we obtain a counterexample to (35).

Example 4. Let $\theta \in [0, 1]$ denote the success probability of a binomial distribution $P_\theta = \text{Bin}(2, \theta)$ on $\mathcal{X} = \{0, 1, 2\}$. Then for $\alpha = \infty$ the redundancy achieving distribution is $S = (\frac{2}{5}, \frac{1}{5}, \frac{2}{5})$ and the minimax redundancy is $R_\infty = \ln \frac{5}{2}$.

In this case there are many barycentric input distributions. For example, the distribution $\pi = \frac{1}{5}M_0 + \frac{3}{5}U + \frac{1}{5}M_1$ is a barycentric input distribution, where M_θ is a point-mass on θ and U is the uniform distribution on $[0, 1]$. Another example is the distribution $\pi = (\frac{3}{10}, \frac{2}{5}, \frac{3}{10})$ on the maximum likelihood parameters $\Psi = \{0, \frac{1}{2}, 1\}$ for the elements of \mathcal{X} .

If there exists a capacity achieving input distribution π_{opt} , then by Theorem 32 it must be such that

$$D_\infty(P_\theta \| S) = \ln \frac{5}{2} \quad (40)$$

almost surely under π_{opt} . The only θ that satisfy (40) are the maximum likelihood parameters in the set Ψ defined above, and hence π_{opt} must be supported on Ψ . Using positivity of Kullback-Leibler divergence (Theorem 8), it can be shown that

the infimum in

$$\begin{aligned} &\inf_Q \sum_{\theta \in \Psi} D_\infty(P_\theta \| Q) \pi_{\text{opt}}(\theta) \\ &= \inf_Q \left\{ \pi_{\text{opt}}(0) \ln \frac{1}{Q(0)} + \pi_{\text{opt}}(\frac{1}{2}) \ln \frac{1/2}{Q(1)} + \pi_{\text{opt}}(1) \ln \frac{1}{Q(2)} \right\} \end{aligned} \quad (41)$$

is uniquely achieved by $Q = (\pi_{\text{opt}}(0), \pi_{\text{opt}}(\frac{1}{2}), \pi_{\text{opt}}(1))$. If π_{opt} is to be the capacity achieving input distribution, then this Q must equal S by Theorem 32, and hence $\pi_{\text{opt}} = (\frac{2}{5}, \frac{1}{5}, \frac{2}{5})$ on Ψ . Evaluating (41) for this choice of π_{opt} , we indeed find that it equals $\ln \frac{5}{2} = R_\infty = C_\infty$ as required, and thus π_{opt} is the unique capacity achieving input distribution.

V. NEGATIVE ORDERS

Until now we have only discussed Rényi divergence of non-negative orders. However, using formula (5) for $\alpha \in (-\infty, 0)$ (reading $\frac{q^{1-\alpha}}{p^{1-\alpha}}$ for $p^\alpha q^{1-\alpha}$), it may also be defined for these negative orders. This definition extends to $\alpha = -\infty$ by

$$D_{-\infty}(P \| Q) = \lim_{\alpha \downarrow -\infty} D_\alpha(P \| Q).$$

According to Rényi [1], only positive orders can be regarded as measures of information, and negative orders indeed seem to be hardly used in applications. Nevertheless, for completeness we will also study Rényi divergence of negative orders. As will be seen below, our results for positive orders carry over to the negative orders, but most properties are reversed. People may have avoided negative orders because of these reversed properties. Avoiding negative orders is always possible, because they are related to orders $\alpha > 1$ by an extension of skew symmetry:

Lemma 12 (Skew Symmetry). *For any $\alpha \in (-\infty, \infty)$, $\alpha \neq \{0, 1\}$*

$$D_\alpha(P \| Q) = \frac{\alpha}{1 - \alpha} D_{1-\alpha}(Q \| P). \quad (42)$$

Furthermore

$$\begin{aligned} D_{-\infty}(P \| Q) &= -D_\infty(Q \| P) \\ &= \ln \inf_{A \in \mathcal{F}} \frac{P(A)}{Q(A)} = \ln \left(\text{ess inf}_Q \frac{p}{q} \right), \end{aligned}$$

with the conventions that $0/0 = 0$ and $x/0 = \infty$ for $x > 0$.

Proof: The identity (42) follows directly from definitions. It implies $D_{-\infty}(P \| Q) = -D_\infty(Q \| P)$, because $\frac{\alpha}{1-\alpha}$ tends to -1 as $\alpha \rightarrow -\infty$. The remaining identities follow from the closed-form expressions for $D_\infty(Q \| P)$ in Theorem 6. ■

Skew symmetry gives a kind of symmetry between the orders $\frac{1}{2} + \alpha$ and $\frac{1}{2} - \alpha$. In applications in physics this symmetry is related to the use of so-called *escort probabilities* [44].

Whereas the nonnegative orders generally satisfy the same or similar properties for different values of α , the fact that $\frac{\alpha}{1-\alpha} < 0$ for $\alpha < 0$, implies that properties for negative orders are often inverted. For example, Rényi divergence for negative orders is nonpositive, concave in its first argument and upper semi-continuous in the topology of setwise convergence.

In addition, the data processing inequality holds with its inequality reversed and for $\alpha \in (-\infty, 0)$ Theorem 2 applies with an infimum instead of a supremum.

Not all properties are inverted, however. Most notably, it does remain true that Rényi divergence is nondecreasing and continuous in α :

Theorem 35. *For $\alpha \in [-\infty, \infty]$, the Rényi divergence $D_\alpha(P\|Q)$ is nondecreasing in α .*

Proof: For $\alpha < 0$, $D_\alpha(P\|Q) \leq 0$ and for $\alpha \geq 0$, $D_\alpha(P\|Q) \geq 0$, so the divergence for negative orders never exceeds the divergence for nonnegative orders. The remainder of the proof follows from Theorem 3 and skew symmetry. ■

Theorem 36. *The Rényi divergence $D_\alpha(P\|Q)$ is continuous in α on*

$$A = \{\alpha \in [-\infty, \infty] \mid 0 \leq \alpha \leq 1 \text{ or } |D_\alpha(P\|Q)| < \infty\}.$$

Proof: Rényi divergence is nondecreasing in α , nonnegative for $\alpha \geq 0$ and nonpositive for $\alpha < 0$. Therefore the required continuity follows directly from Theorem 7 and skew symmetry, except for the case

$$\lim_{\alpha \uparrow 0} D_\alpha(P\|Q) = D_0(P\|Q),$$

which is required to hold if there exists a value $\beta < 0$ such that $D_\beta(P\|Q) > -\infty$. In this case $D_{1-\beta}(Q\|P) = \frac{1-\beta}{\beta} D_\beta(P\|Q) < \infty$, which implies: (a) that $Q \ll P$, so $D_0(P\|Q) = 0$; and (b) that $D(Q\|P) < \infty$ and by Theorem 5

$$\lim_{\alpha \uparrow 0} D_\alpha(P\|Q) = \lim_{\alpha \uparrow 0} \frac{\alpha}{1-\alpha} D_{1-\alpha}(Q\|P) = 0 \cdot D(Q\|P) = 0.$$

■

VI. SUMMARY

We have reviewed and derived the most important properties of Rényi divergence and Kullback-Leibler divergence. These include convexity and continuity properties, limits of σ -algebras, additivity for product distributions on infinite sequences, and the relation of the special order 0 to absolute continuity and mutual singularity of such distributions.

We have also derived several key minimax identities. In particular, Theorems 27 and 28 illuminate the relation between Rényi divergence, Kullback-Leibler divergence and Chernoff information in hypothesis testing. And Theorem 30 extends the known equivalence of channel capacity and minimax redundancy (for all orders) to continuous channel inputs. A new result relating the worst-case redundancy of a distribution to its divergence from the (unique) minimax redundancy achieving distribution was given by Theorem 34.

ACKNOWLEDGMENTS

The authors would like to thank Peter Grünwald and Wouter Koolen for useful discussions. Part of the work was done while both authors were with the Centrum Wiskunde & Informatica in Amsterdam, the Netherlands, and while Tim van Erven was with the VU University, also in Amsterdam.

APPENDIX: NEGATIVE RESULTS

Some useful properties that are satisfied by other divergences, are not satisfied by Rényi divergence. Here we give counterexamples for a few important ones.

A. No Pythagorean Inequality

An important result in statistical applications of information theory is the Pythagorean inequality for Kullback-Leibler divergence [27], [45], [46]. It states that, if \mathcal{P} is a convex set of distributions, Q is any distribution not in \mathcal{P} , and $D_{\min} = \inf_{P \in \mathcal{P}} D(P\|Q)$, then there exists a distribution P^* such that

$$D(P\|Q) \geq D(P\|P^*) + D_{\min} \quad \text{for all } P \in \mathcal{P}.$$

The main use of the Pythagorean inequality lies in its implication that if P_1, P_2, \dots is a sequence of distributions in \mathcal{P} such that $D(P_n\|Q) \rightarrow D_{\min}$, then P_n converges to P^* in the strong sense that $D(P_n\|P^*) \rightarrow 0$.

Unfortunately, for $\alpha \neq 1$ Rényi divergence does not satisfy the Pythagorean inequality, as demonstrated by the counterexamples below. We should point to results by Sundaresan [47], however, who argues that, under regularity conditions, for finite sample spaces a generalisation of Rényi divergence (see [48]) does satisfy a modified Pythagorean inequality, in which every distribution $R \in \{P, Q\}$ is replaced by its *tilted* counterpart

$$R'(x) = \frac{R(x)^\alpha}{\sum_y R(y)^\alpha}.$$

To construct the counterexamples for the ordinary Pythagorean inequality, first consider $\alpha \in [0, 1)$. Let $Q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ be uniform on three points and let $\mathcal{P} = \{(p_1, p_2, p_3) \mid p_1 = \frac{1}{4}\}$ be the convex set of distributions with first component fixed at $\frac{1}{4}$. Then $\inf_{P \in \mathcal{P}} D_\alpha(P\|Q)$ is achieved by $P^* = (\frac{1}{4}, \frac{3}{8}, \frac{3}{8})$ and the Pythagorean inequality

$$D_\alpha(P\|Q) \geq D_\alpha(P\|P^*) + D_\alpha(P^*\|Q) \quad (43)$$

is violated for $P = (\frac{1}{4}, 0, \frac{3}{4})$: if $\alpha > 0$, then (43) is equivalent to

$$1 + 3^\alpha \leq \left(\frac{1}{4} + \frac{3}{8} 2^\alpha\right) \left(1 + 2 \left(\frac{3}{2}\right)^\alpha\right) \\ (1 - 2^{1-\alpha})(23^\alpha - 32^\alpha) \leq 0,$$

which is false. If $\alpha = 0$, then $D_\alpha(P\|Q) = -\ln \frac{2}{3}$, $D_\alpha(P\|P^*) = -\ln \frac{5}{8}$ and $D_\alpha(P^*\|Q) = 0$, and the inequality does not hold either.

Secondly, for $\alpha \in (1, \infty]$ take $Q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $\mathcal{P} = \{(p_1, p_2, p_3) \mid p_1 = \frac{2}{3}\}$. Then $\inf_{P \in \mathcal{P}} D_\alpha(P\|Q)$ is achieved by $P^* = (\frac{2}{3}, \frac{1}{6}, \frac{1}{6})$ and the inequality is violated for $P = (\frac{2}{3}, 0, \frac{1}{3})$: if $\alpha < \infty$, then (43) is equivalent to

$$6(1 + 2^\alpha) \geq (4 + 2^\alpha)(2^{1-\alpha} + 2^\alpha) \\ (2^\alpha - 2)(4^\alpha - 4) \leq 0,$$

which is false. If $\alpha = \infty$, then $D_\alpha(P\|Q) = D_\alpha(P\|P^*) = D_\alpha(P^*\|Q) = \ln 2$ and the inequality does not hold either.

B. Convexity in P does not hold for $\alpha > 1$

Rényi divergence for $\alpha \in (1, \infty)$ is not convex in its first argument. Consider the following counterexample: let $0 < p_0 < p_1 < 1$ be any two numbers, and let $p_{1/2} = \frac{p_0 + p_1}{2}$. Let $\varepsilon > 0$ be arbitrary, and let $0 < q < 1$ be small enough that

$$\max_{i \in \{0,1\}} \frac{(1 - p_i)^\alpha (1 - q)^{1-\alpha}}{p_i^\alpha q^{1-\alpha}} \leq \varepsilon.$$

Then convexity of D_α in its first argument would imply that

$$\begin{aligned} & \frac{1}{2} \ln (p_0^\alpha q^{1-\alpha} + (1 - p_0)^\alpha (1 - q)^{1-\alpha}) \\ & + \frac{1}{2} \ln (p_1^\alpha q^{1-\alpha} + (1 - p_1)^\alpha (1 - q)^{1-\alpha}) \\ & \geq \ln (p_{1/2}^\alpha q^{1-\alpha} + (1 - p_{1/2})^\alpha (1 - q)^{1-\alpha}), \end{aligned}$$

which implies

$$\begin{aligned} & \frac{1}{2} \ln (p_0^\alpha q^{1-\alpha} (1 + \varepsilon)) + \frac{1}{2} \ln (p_1^\alpha q^{1-\alpha} (1 + \varepsilon)) \geq \ln (p_{1/2}^\alpha q^{1-\alpha}) \\ & \frac{1}{2} \ln (p_0^\alpha (1 + \varepsilon)) + \frac{1}{2} \ln (p_1^\alpha (1 + \varepsilon)) \geq \ln (p_{1/2}^\alpha). \end{aligned}$$

As this expression holds for all $\varepsilon > 0$, we get

$$\begin{aligned} & \frac{1}{2} \ln p_0^\alpha + \frac{1}{2} \ln p_1^\alpha \geq \ln p_{1/2}^\alpha \\ & \frac{1}{2} \ln p_0 + \frac{1}{2} \ln p_1 \geq \ln \frac{p_0 + p_1}{2}, \end{aligned}$$

which is a contradiction, because the natural logarithm is strictly concave.

C. Rényi divergence is not continuous

In general the Rényi divergence of order $\alpha \in (0, 1)$ is not continuous in the topology of setwise convergence. To construct a counterexample, recall that $\tau = 2\pi$, let P_n denote the probability distribution on $[0, \tau]$ with density $\frac{1 + \sin(nx)}{\tau}$ and let Q_n denote the probability distribution on $[0, \tau]$ with density $\frac{1 - \sin(nx)}{\tau}$ for $n = 1, 2, \dots$. Then $D_\alpha(P_n \| Q_n) > 0$ does not depend on n , and both P_n and Q_n converge to the uniform distribution U on $[0, \tau]$ in the topology of setwise convergence. Consequently, $\lim_{n \rightarrow \infty} D_\alpha(P_n \| Q_n) \neq 0 = D_\alpha(U \| U)$, so in general D_α is not continuous in the topology of setwise convergence.

D. Not a metric

Except for the order $\alpha = \frac{1}{2}$, Rényi divergence is not symmetric and cannot be a metric. For $\alpha = \frac{1}{2}$ Rényi divergence is symmetric and by (2) it locally behaves like the square of a metric. Therefore one may wonder whether it actually is the square of a metric itself. Consider the following three distributions on two points:

$$P = (0, 1), \quad Q = \left(\frac{1}{2}, \frac{1}{2}\right), \quad R = (1, 0).$$

Then

$$D_{\frac{1}{2}}(P \| Q) = \ln 2, \quad D_{\frac{1}{2}}(Q \| R) = \ln 2, \quad D_{\frac{1}{2}}(P \| R) = \infty.$$

As the square roots of these divergences violate the triangle inequality, $D_{1/2}$ cannot be the square of a metric.

REFERENCES

- [1] A. Rényi, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 547–561, 1961.
- [2] P. Harremoës, "Interpretations of Rényi entropies and divergences," *Physica A: Statistical Mechanics and its Applications*, vol. 365, no. 1, pp. 57–62, 2006.
- [3] P. D. Grünwald, *The Minimum Description Length Principle*. The MIT Press, 2007.
- [4] I. Csiszár, "Generalized cutoff rates and Rényi's information measures," *IEEE Transactions on Information Theory*, vol. 41, no. 1, pp. 26–34, 1995.
- [5] T. Zhang, "From ϵ -entropy to KL-entropy: Analysis of minimum information complexity density estimation," *The Annals of Statistics*, vol. 34, no. 5, pp. 2180–2210, 2006.
- [6] D. Haussler and M. Opper, "Mutual information, metric entropy and cumulative relative entropy risk," *The Annals of Statistics*, vol. 25, no. 6, pp. 2451–2492, 1997.
- [7] L. Le Cam, "Convergence of estimates under dimensionality restrictions," *The Annals of Statistics*, vol. 1, no. 1, pp. 38–53, 1973.
- [8] L. Birgé, "On estimating a density using Hellinger distance and some other strange facts," *Probability Theory and Related Fields*, vol. 71, pp. 271–291, 1986.
- [9] S. van de Geer, "Hellinger-consistency of certain nonparametric maximum likelihood estimators," *The Annals of Statistics*, vol. 21, no. 1, pp. 14–44, 1993.
- [10] D. Morales, L. Pardo, and I. Vajda, "Rényi statistics in directed families of exponential experiments," *Statistics*, vol. 34, pp. 151–174, 2000.
- [11] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Multiple source adaptation and the Rényi divergence," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 367–374, 2009.
- [12] A. O. Hero, B. Ma, O. Michel, and J. D. Gorman, "Alpha-divergence for classification, indexing and retrieval (revised)," Tech. Rep. CSPL-334, Communications and Signal Processing Laboratory, The University of Michigan, 2003.
- [13] J. Aczél and Z. Daróczy, *On Measures of Information and Their Characterizations*. Academic Press, 1975.
- [14] M. Ben-Bassat and J. Raviv, "Rényi's entropy and the probability of error," *IEEE Transactions on Information Theory*, vol. 24, no. 3, pp. 324–330, 1978.
- [15] T. van Erven and P. Harremoës, "Rényi divergence and majorization," in *IEEE International Symposium on Information Theory (ISIT)*, 2010.
- [16] T. van Erven, *When Data Compression and Statistics Disagree: Two Frequentist Challenges for the Minimum Description Length Principle*. PhD thesis, Leiden University, 2010.
- [17] O. Shayelevitz, "A note on a characterization of Rényi measures and its relation to composite hypothesis testing," arXiv:1012.4401v1, Dec. 2010.
- [18] O. Shayelevitz, "On Rényi measures and hypothesis testing," in *IEEE International Symposium on Information Theory Proceedings*, pp. 800–804, 2011.
- [19] F. Liese and I. Vajda, *Convex Statistical Distances*. Leipzig: Teubner, 1987.
- [20] M. Gil, "On Rényi divergence measures for continuous alphabet sources," Master's thesis, Queen's University, 2011.
- [21] D. Aldous and P. Diaconis, "Strong uniform times and finite random walks," *Advances in Applied Mathematics*, vol. 8, pp. 69–97, 1987.
- [22] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *International Statistical Review*, vol. 70, pp. 419–435, 2002.
- [23] G. L. Gilardoni, "On Pinsker's and Vajda's type inequalities for Csiszár's f -divergences," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5377–5386, 2010.
- [24] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4394–4412, 2006.
- [25] A. N. Shiryaev, *Probability*. Springer-Verlag, 1996.
- [26] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society, series B*, vol. 28, no. 1, pp. 131–142, 1966.
- [27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, 1991.
- [28] O. Kallenberg, *Foundations of Modern Probability*. Springer, 1997.
- [29] J. Feldman, "Equivalence and perpendicularity of Gaussian processes," *Pacific Journal of Mathematics*, vol. 8, no. 4, pp. 699–708, 1958.

- [30] J. Hájek, "On a property of normal distributions of any stochastic process," *Czechoslovak Mathematical Journal*, vol. 8, no. 4, pp. 610–618, 1958. In Russian with English summary.
- [31] B. J. Thelen, "Fisher information and dichotomies in equivalence/contiguity," *The Annals of Probability*, vol. 17, no. 4, pp. 1664–1690, 1989.
- [32] S. Kakutani, "On equivalence of infinite product measures," *The Annals of Mathematics*, vol. 49, no. 1, pp. 214–224, 1948.
- [33] A. Rényi, "On some basic problems of statistics from the point of view of information theory," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1: Statistics, pp. 531–543, 1967.
- [34] A. W. van der Vaart, *Asymptotic Statistics*. Cambridge University Press, 1998.
- [35] T. Nemetz, "On the α -divergence rate for Markov-dependent hypotheses," *Problems of Control and Information Theory*, vol. 3, no. 2, pp. 147–155, 1974.
- [36] Z. Rached, F. Alajaji, and L. L. Campbell, "Rényi's divergence and entropy rates for finite alphabet Markov sources," *IEEE Transactions on Information Theory*, vol. 47, no. 4, pp. 1553–1561, 2001.
- [37] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, 1970.
- [38] B. Ryabko, "Comments on "a source matching approach to finding minimax codes" by Davisson, L. D. and Leon-Garcia, A.," *IEEE Transactions on Information Theory*, vol. 27, no. 6, pp. 780–781, 1981. Including also the ensuing Editor's Note.
- [39] D. Haussler, "A general minimax result for relative entropy," *IEEE Transactions on Information Theory*, vol. 43, no. 4, pp. 1276–1280, 1997.
- [40] M. Sion, "On general minimax theorems," *Pacific Journal of Mathematics*, vol. 8, no. 1, pp. 171–176, 1958.
- [41] H. Komiya, "Elementary proof for Sion's minimax theorem," *Kodai Mathematical Journal*, vol. 11, no. 1, pp. 5–7, 1988.
- [42] Y. M. Shtar'kov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, no. 3, pp. 175–186, 1987.
- [43] R. Sibson, "Information radius," *Z. Wahrscheinlichkeitstheorie verw. Geb.*, vol. 14, pp. 149–160, 1969.
- [44] J. Naudts, "Estimators, escort probabilities, and ϕ -exponential families in statistical physics," *Journal of Inequalities in Pure and Applied Mathematics*, vol. 5, no. 4, 102, 2004.
- [45] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *The Annals of Probability*, vol. 3, no. 1, pp. 146–158, 1975.
- [46] F. Topsøe, *Entropy, Search, Complexity*, vol. 16 of *Bolyai Society Mathematical Studies*, ch. 8, Information Theory at the Service of Science, pp. 179–207. Springer, 2007.
- [47] R. Sundaresan, "A measure of discrimination and its geometric properties," in *IEEE International Symposium on Information Theory (ISIT)*, 2002.
- [48] R. Sundaresan, "Guessing under source uncertainty with side information," in *IEEE International Symposium on Information Theory (ISIT)*, 2006.